THE IMPACT OF SCIENCE NOTEBOOK WRITING ON ELL AND LOW-SES STUDENTS' SCIENCE LANGUAGE DEVELOPMENT AND CONCEPTUAL UNDERSTANDING

A Dissertation

by

MARGARITA HUERTA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Rafael Lara-Alecio
Co-Chair of Committee,	Fuhui Tong
Committee Members,	Sharolyn Pollard-Durodola
	Yeping Li
	Beverly Irby
Head of Department,	Victor Willson

May 2013

Major Subject: Educational Psychology

ABSTRACT

This quantitative study explored the impact of literacy integration in a science inquiry classroom involving the use of science notebooks on the academic language development and conceptual understanding of students from diverse (i.e., English Language Learners, or ELLs) and low socio-economic status (low-SES) backgrounds. The study derived from a randomized, longitudinal, field-based NSF funded research project (NSF Award No. DRL - 0822343) targeting ELL and non-ELL students from low-SES backgrounds in a large urban school district in Southeast Texas. The study used a scoring rubric (modified and tested for validity and reliability) to analyze fifth-grade school students' science notebook entries.

Scores for academic language quality (or, for brevity, *language*) were used to compare language growth over time across three time points (i.e., beginning, middle, and end of the school year) and to compare students across categories (ELL, former ELL, non-ELL, and gender) using descriptive statistics and mixed between-within subjects analysis of variance (ANOVA). Scores for conceptual understanding (or, for brevity, *concept*) were used to compare students across categories (ELL, former ELL, non-ELL, and gender) in three domains using descriptive statistics and ANOVA. A correlational analysis was conducted to explore the relationship, if any, between language scores and concept scores for each group.

Students demonstrated statistically significant growth over time in their academic language as reflected by science notebook scores. While ELL students scored lower than former ELL and non-ELL students at the first two time points, they caught up to their peers by the third time point. Similarly, females outperformed males in language scores in the first two time points, but males caught up to females in the third time point. In analyzing conceptual scores, ELLs had statistically significant lower scores than former-ELL and non-ELL students, and females outperformed males in the first two domains. These differences, however, were not statistically significant in the last domain. Last, correlations between language and concept scores were overall, positive, large, and significant across domains and groups. The study presents a rubric useful for quantifying diverse students' science notebook entries, and findings add to the sparse research on the impact of writing in diverse students' language development and conceptual understanding in science.

DEDICATION

To my family, for their unconditional love and support

ACKNOWLEDGEMENTS

I acknowledge gratitude to the incredible professors at Texas A&M: Dr. Lara, my chair, for the opportunity and generous support; Dr. Tong, my co-chair, for insight into the analysis and results; Dr. Irby for aiding in the conceptualization of the project; Dr. Durodola, for feedback on my writing; and Dr. Li for the support and encouragement. Thanks also to Dr. Padrón for believing in my ability and to Dr. Goodson, my unofficial and gracious mentor – without you I would not have completed the writing process.

I further acknowledge gratitude to my friends in the Bilingual Education Program: Tiberio, for always making me laugh and being my constant supporter and prayer-warrior; Janie, for holding us all together; Polly, for being an example of empathy and strength; Julia, for bringing that craziness we sometimes need; Gaby, for being there to talk and listen; Yinan, for sharing laughter; Raymond, for listening and caring; Seth, for the encouraging words and laughs. Thanks to the EPSY staff: Cathy, for coming to visit us with words of kindness; Kristie for being so efficient and kind through the process. Gratitude also to the POWER Writing Group, with whom I've grown to see myself as a professional writer.

My heartfelt gratitude to my family: My dad, for being consistent and committed; my mom, for being my pillar of strength; Rodrigo, for being an example of courage; Rebekah, for being my friend; David, for always loving me.

My eternal gratitude to my Lord and Savior Jesus Christ without whom I would not have the breath of life, to whom all glory belongs, and with whom life has meaning. *Of making many books there is no end, and much study wearies the body. Now all has been heard; here is the conclusion of the matter: Fear God and keep his commandments, for this is the duty of all mankind. -- Ecclesiastes 12:12b -13*

TABLE OF CONTENTS

Page

CHAPTER Page

Page

CHAPTER Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Students who come to school speaking a language other than English and must acquire English language proficiency (Cloud, Genesee, & Hamayan, 2009) are a large and growing part of the educational system in the United States. Researchers refer to this group of students as English Language Learners (ELLs). ELLs make up approximately 10% of the public school population in the United States and are projected to continue to increase in number (National Center for Education Statistics [NCES], 2012). In Texas, the percentage of students receiving ELL or bilingual services increased by 56.4 percent from 2000 to 2011 and is projected to continue to increase (Texas Education Agency [TEA], 2011a).

Ethnically diverse or non-mainstream students, or "…students of color, students learning English as a new language, students from immigrant or low-income families," who have cultural and linguistic backgrounds different from white, mainstream American culture (Lee & Luykx, 2005. p. 413) are also a large and growing part of the public education system in the United States. Hispanic students currently make up 23% of the public school population in the United States (NCES, 2012). In Texas, the enrollment of African American and Hispanic students increased from 2009 to 2011, while the enrollment of White students decreased – currently, Hispanic students account for 50.3% of the student enrollment; White students account for 31.2%, and African American students account for 12.9% of the public school enrollment (TEA, 2011a). In addition, the percentage of low-SES students in Texas increased by 45.5 % from 2000 to 2011 – more than double the 21.5 % increase of low-SES students in the nation as a whole (TEA, 2011a). In the United States, approximately 19% of school age children were enrolled in high-poverty schools, where 76% or more of the students qualify for free or reduced lunch due to low-SES backgrounds (NCES, 2012). In Texas, 77.4% of Hispanic students and 71.6% of African American Students classify as low-SES (TEA, 2011a).

Individuals in a society need to be well educated in science, especially in light of the exponential growth of scientific and technological innovation and the need for science knowledge in a global economy (National Science Board, 2010). The *National Science Education Standards* (National Research Council, 1996) define being educated in science as being *scientifically literate*, which includes being able to understand scientific concepts and to use language to "describe*,* explain, and predict natural phenomenon" (p. 22). The need for students to be scientifically literate is therefore critical if the United States is to foster a society able to make educated decisions related to science.

Educators no longer just face the challenge of facilitating English proficiency for ELL students and academic language for ELL and low-SES students, but of facilitating the acquisition of academic language and content-area understanding simultaneously (i.e., part of what it means to be scientifically literate). The *National Science Standards* include the phrase, "Science is for all students," as one of its guiding principals (National Research Council, 1996, p. 19). Researchers, however, have noted that ELL and low-SES students have difficulty in learning content-area subjects due to linguistic and cultural differences between their home culture and the culture of the school (Lee, 2005; Lee & Luykx, 2005).

The challenge to build scientific literacy for ELL and low-SES students is evident across grade levels as reflected in national science achievement scores (NCES, 2011). The challenge, however, becomes greater as students move into the upper elementary and middle grades where the demands of scientific language and content increase (Fang, 2006; Merino & Scarcella, 2005) while ELL and low-SES students continue to fall behind in their academic achievement in the content areas such as science (NCES, 2011).

Definition of Terms

ELL

English Language Learner; a student who is in the process of learning English as a second or other language and may benefit from various types of language support programs (National Council of Teachers of English, 2008).

Former ELL

Former English Language Learner; a student who was previously classified as ELL but has been exited from a bilingual or English as a Second Language (ESL program). In the state in which the present study took place, ELLs are exited from a bilingual or ESL program based on the decision of a review committee comprised of parent representative(s), teacher(s), administrator(s), and any educational specialists. The committee annually reviews students' progress based on the following criteria: (a) state approved tests that measure the extent to which the student has developed oral and written language proficiency and specific language skills in English; (b) satisfactory performance on state approved reading assessment instruments; (c) state-approved criterion-referenced written tests if available, and the results of subjective teacher evaluation (TEA, 2011-2012).

Low-SES

Low socioeconomic status; a category often identified by whether a student qualifies for free or reduced lunch; the majority of nonmainstream students classify as low-SES (NCES, 2011). In this study, low-SES is synonymous to the term "non-ELL" because all of the students in the study sample were from low-SES backgrounds.

Academic Language

Language used in the learning of academic subject matter in a formal schooling context; aspects of language strongly associated with literacy and academic achievement, including specific academic terms or technical language, and speech registers related to each field of study (Teachers of English to Speakers of Other Languages, 1997).

Conceptual Understanding

What it is not: Science knowledge refers to facts, concepts, principles, laws, theories, and models. What it is: Science conceptual understanding is the ability to use science knowledge (National Research Council, 1996).

Scientific Literacy

Scientific literacy is a complex term including the overarching idea that individuals should be able to know and understand science in order to make informed decisions in society as well as apply specialized skills which involve speaking, reading, writing, and thinking about science (National Research Council, 1996).

At its core, scientific literacy encompasses the idea that science constitutes a specialized language. This language includes academic language (Gee, 2005) and thinking skills needed in order to understand science concepts (Merino & Scarcella, 2005), learn science concepts (Halliday & Martin, 1993), and express science concepts (Rivard & Straw, 2004).

Science Inquiry

An instructional model based on the constructivist idea that individuals learn by making connections between new information and prior knowledge (Rosebery, Warren, & Conant, 1992). One structured model of science inquiry is the 5-E model developed by Bybee et al. (1996) in which teachers guide students in the process of engagement, exploration, explanation, elaboration, and evaluation within one science lesson.

Science Notebook

A tool used in the classroom as a space for students to write down scientific questions, investigations, procedures, reflections, and conclusions (Butler & Nesbit, 2008).

Rubric

In education, a rubric is a set of criteria on a continuum, meant to describe varying levels of performance on a given task (Luft, 1998).

Analysis of Variance

A statistical analysis used to evaluate the equality of means on a single outcome variable that is at least intervally-scaled, across two or more groups (Thompson, 2008).

Repeated Measures Analysis of Variance

A statistical analysis that is used to measure subjects on the same continuous scale on three or more occasions (Pallant, 2010).

Mixed Between-Within Subjects Analysis of Variance

A statistical technique that is an extension of repeated measures analysis of variance; it allows combining between-subjects and within-subjects in one analysis (Pallant, 2010).

Correlational Analysis

Correlational analysis is a bivariate statistical analysis used to describe the direction and strength (and existence if any) of the linear relationship between two variables (Pallant, 2010; Thompson, 2008).

Statement of the Problem

Scientific literacy, the ability to use science language to understand, learn, and express science concepts, is critical for ELL and low-SES students to achieve academically in science (Kieffer, Lesaux, Rivera, and Francis, 2009; Lee & Lyukx, 2005). Yet, ELL and low-SES students continue to fall behind in science achievement at a national (NCES, 2011) and state level (TEA, 2011).

For example, according to NCES (2011) in 4th grade, only 47% of African American and 53% of Hispanic students scored at or above the basic science achievement level compared to 87% of White students. Moreover, only 11% of African American and 14% of Hispanic students reached proficient science levels as compared to 47% of White students. In 8th grade, the gap increased by about 10% at the basic level and remained comparable at the proficient level between minority and mainstream groups.

Data from the state where the study took place had similar trends. In 5th grade, 79% of African American students, 83% of Hispanic, 71% of ELL students, and 94% of White students (TEA, 2011b) met the state science test standards. In 8th grade, the gap between nonmainstream and mainstream students increased: 69% of African American students, 73% of Hispanic, and 44% of ELL students met the science test standards while 90% of White students met the science test standards (TEA, 2011b).

Students classified as low-SES also fall behind on science achievement at national (NCES, 2011), and state (TEA, 2011b) levels. In 4th grade science achievement tests, 54% of students eligible for free lunch, an indicator for low-SES, compared to 86% of students not eligible for free lunch scored at or above the basic level in science achievement. The gap increased at the proficient levels in science achievement: only 15% of students eligible for free lunch reached this level, compared with 48% of students not eligible for free lunch. In 8th grade, the gap remained the same (about a 10% and 30% difference) (NCES 2011).

Data from the state in which the study was conducted allows researchers to differentiate between an *economically disadvantaged* (i.e., low-SES) passing rate and a non-economically disadvantaged passing rate, but the data does not include a category excluding low-SES students. It is therefore impossible to compare achievement gaps between low SES and non low-SES students. It is possible, however, to know that in 5th grade, 82% of low-SES students met the state standards for science while 71% of low SES 8th graders met the standard (TEA, 2011b).

Theoretical Framework

Researchers have attributed the achievement gaps in science between ELL and low-SES as compared to mainstream students to two main factors. The first is a lack of *academic language* needed in order to understand science concepts in English and succeed on standardized assessments (Kieffer et al., 2009; Lyukx et al., 2007; Wolf & Leon, 2009). The second is a lack of student *conceptual understanding* of science needed to participate in science learning (Lee & Fradd, 1996; Hudicourt-Barnes, 2003; Lee, 2005). The factors will be discussed below from a theoretical perspective. Each section ends with a paragraph noting the critical application of the theory to educational practice.

Academic Language

Academic language has been defined as aspects of language strongly associated with literacy and academic achievement, including specific academic terms or technical language, and speech registers related to each field of study (Teachers of English to Speakers of Other Languages, 1997). Language is so critical to learning that theorist in the field of ELL (e.g., Cummins, 1981; Scarcella, 2003) and science education (e.g., Gee, 2005; Halliday & Martin, 1993) have defined types of language in academic settings as will be discussed below.

BICS vs. CALP. Cummins (1981) made a distinction between Basic Interpersonal Communication (BICS) and Cognitive Academic Language Proficiency (CALP) when discussing ELLs' learning of language in academic settings. BICS is the social language L2 learners acquire that is considered cognitively undemanding. CALP, on the other hand, is the *academic language* L2 learners must acquire in their second language in order to be successful in school settings. Cummins (1981) noted that if ELLs only acquired BICS, they would not be able to succeed in academic settings. More recently, Scarcella (2003) noted, "Learning academic English is probably one of the surest, most reliable ways of attaining socio-economic success in the United States today. Learners cannot function in school settings effectively without it" (p. 3). Scarcella (2003) elaborated how academic English entails, "mastery of a writing system and its particular academic conventions as well as proficiency in reading, speaking, and listening" (p. 3). Academic language – in all its forms – is inevitably critical for ELL's success in academia and beyond the classroom.

Academic vs. everyday English. Scarcella (2003), however, criticized BICS and CALP for being simplistic and dichotomous. She argued that BICS or “everyday” language and CALP or “academic English” share related components, making the boundaries between BICS and CALP “fuzzy” (p. 27). In specific, Scarcella (2003) noted *linguistic components* needed for both everyday and academic English (i.e., phonological, lexical, grammatical, sociolinguistic, and discourse components). The components are important building blocks for everyday and academic English as well as blocks that may or may not be dependent on each other (i.e., students may acquire a simple academic vocabulary word before acquiring a specific phonological concept). Scarcella (2003) further noted *cognitive components* are present in both everyday and academic English (i.e., knowledge, higher order thinking, strategic, and metalinguistic components).

Nonetheless, Scarcella (2003) emphasizes the difference between everyday and academic English by illustrating linguistic and cognitive features that become more important in academic written discourse (e.g., higher order thinking, grammatical features, background knowledge, meta-linguistic abilities) vs. everyday conversation (e.g., phonological features, discourse features, sociolinguistic features). Arguably, the features characteristic in “everyday conversation” could be used to describe more informal writing, and, Scarcella (2003) does this in comparing a creative informal poem to a more academic expository piece of writing. Such specific distinctions are helpful in thinking about how to classify and measure academic language and perhaps even conceptual understanding.

Scientific vs. everyday language. Like Cummins (1981) and Scarcella (2003), theorists in science education have made a distinction between “everyday” language (like BICS) and the “scientific language” (like CALP or academic English) (Gee, 2005; Halliday & Martin, 1993; Lemke, 1990). According to the theorists, science language is made up of distinctive linguistic features such as technical vocabulary as well as specific discourse patterns that differ from the everyday language that students use outside of the science classroom (Gee, 2005; Lemke, 1990). For example, Lemke (1990) classified science language into two major *genres.* To clarify, Scarcella (2003) explains genres as “discourse types” with complete structures and identifiable formal properties and purposes. Lemke’s (1990) minor science genres include short descriptions, comparisons, and definitions. Major science genres include longer pieces such as laboratory reports.

Like Cummins (1981) and Scarcella (2003), science theorists have acknowledged the importance of academic language to learning. Halliday and Martin (1993), for example, noted how even speakers whose first language (L1) is English must recognize academic language in science as a type of English (Halliday & Martin, 1993). Gee (2005) and Lemke (1990) further noted how students need this academic language in order to be able to engage in scientific discourse (i.e., listening, speaking, reading and writing about science).

Academic language in the classroom. Researchers in the realm of bilingual education have noted that academic English takes longer to learn in the classroom (approximately seven years) than social English (approximately three years) (Collier & Thomas, 1989). As Scarcella (2003) pointed out, however, students develop academic

English only if they are exposed to it, which may or may not happen in their local communities or even in their schools. Researchers should therefore measure the academic language development of former ELLs (i.e., students who were once classified as ELL and then were exited from ELL programs) and non-ELLs (i.e., native English speakers who may live in communities where they have no exposure to academic English) in order to gage their exposure and successful acquisition of academic English in the classroom over time in addition, of course, to critically tracking ELL development.

Whether individuals possess academic language or not holds serious implications for science learning and instruction for all students. Without academic language, students cannot comprehend content being delivered (e.g., speaking), consumed (e.g., reading) and cannot produce content (e.g., speaking, writing). This fact makes academic language a barrier to many students' science learning (Ryoo, 2010) *as well as* an important component of instruction in order for students to succeed in science. The importance of academic language appears especially true for ELL (Cummins, 1981; Scarcella, 2003) and low-SES students (Norris & Phillips, 2003) based on research analysis implying strong connections between science achievement and academic language (Kieffer et al., 2009).

Conceptual Understanding

Entwistle (2007) defined a concept as a grouping of object, behaviors, or ideas. Conceptual understanding is the acquisition of concepts about a particular topic or idea.

The following discusses theory on conceptual understanding as well as the theory that has most influenced science education research and practice.

Dueling theories. Cognitive and constructivist learning theories are important for understanding how individual acquire concepts. Both theories are based on the idea that concepts are classifications of ideas in the mind. However, *how* individuals acquire conceptual understanding is the point of difference between the theories.

On the one hand, cognitive learning theorists view individuals as acquiring conceptual understanding as they gather concepts that are clearly defined and differentiated from one another (Ausubel, Novak, & Hanesian, 1978; Baddeley, 1976). In this sense, learning new concepts means replacing old concepts; so in academic learning, for example, the teacher seeks to replace students' misconceptions about the world with accurate conceptions about the world. On the other hand, constructivist learning theorists view the mind as a place where concepts are formed based on individuals' contextualized experiences (Halldén, 1999; Kelly, 1955). In this sense, misconceptions are not replaced, but rather restructured to more closely reflect an accurate conception within a given context.

Constructivist theory and science education. Science education theory and practice, stemming from the National Science Standards (National Research Council, 1996) is based on constructivist theory and teaching practices that promote constructivist learning (i.e., science inquiry). The following will therefore elaborate on current ideas behind constructivist theory.

In constructivist theory, learning is defined as a process by which individuals construct knowledge by making connections between new information and prior (i.e., background) knowledge. Halldén (1999) noted that the restructuring of concepts (i.e., conceptual understanding) can happen in different contexts, leading to differing results in understanding (i.e., naïve vs. expert understanding). Halldén (1999) explained three contexts in which individuals can acquire conceptual understanding: situational, cognitive, and cultural. Situational contexts are, for example, everyday experiences individuals have with parents or peers and may lead to naïve conceptual understandings. The idea is very much like Cummins' (1981) BICS and science language theorists' "everyday language". Cognitive contexts on the other hand are experiences in, for example, educational institutions where institutions tend to hold generally agreed concepts. Theoretically, the individual in a cognitive context could build more accurate or scientific concepts. Again, parallels can be drawn to the idea of Cummins' (1981) CALP, Scarcella's academic English, and to science theorists' idea of "scientific language". As Halldén (1999) explained, however, students in academic contexts require multiple exposures to ideas in order to create links between their current and growing knowledge and the concepts they are yet to understand. Finally, Halldén (1999) noted that cultural context, or the form of discourse in which the concept is discussed (i.e., everyday language or academic language), affects an individuals' overall understanding of the world based on multiple exposures to experiences. In a sense, the idea of cultural context overarches the idea of whether the individual is in a situational or cognitive context and how that context influences their understanding. Once again, parallels to

ELL and science education theorists ideas of “everyday”/BICS vs. “scientific”/CALP discourse and language can be drawn.

Conceptual understanding in the classroom. Based on constructivist theory, it follows that the National Science Standards (National Research Council, 1996) emphasizes science inquiry instruction in which students are expected to build their own knowledge as teachers facilitate and encourage students to ask questions, hypothesize, experiment, and draw inferences from science experiences and experiments in the classroom (Rosebery, Warren, & Conant, 1992). In this sense, students exposed to repeated experiences in order to build their understanding of science while also being directed through the instructors’ guidance (i.e., scaffolding) towards “accurate” conceptions. Based on the latter idea, researchers have advocated integrating literacy into science instruction, acknowledging the power of discourse on building conceptual understanding in science education (e.g., Fang, 2004, 2006; Lee, 2005). In specific, researchers advocate the idea that academic language possesses the discourse structures students need in order to build, not just science knowledge facts, but scientific understanding which is academic language applied in speaking, listening, reading, and writing (National Research Council, 1996).

Relationship between Language and Concept

The relationship between language and concept is a complex idea and one that has been explored and debated in linguistic theory (i.e., Do we need language to think and understand the world? Or, can individuals think apart from language? Does language shape our understandings of the world?). The following does not attempt to unravel debates on the ideas in linguistic theory (readers interested in the ideas are directed to Pinker, 1994; Vygotsky, 1962), but rather to contextualize the ideas to their role in shaping theory related to how ELL students learn as well as to how ELL and low-SES students learn in the context of the science classroom.

ELL context. Cummins' (1986) illustrated how language and conceptual understanding are related. He used his theory of transference (Cummins, 1986) to explain how L2 learners posses a Common Underlying Proficiency, or CUP, made up of concepts in the mind. These concepts can be transferred to the surface level features of the L2. In this way, L1 and L2 interact with each other. In simpler terms, languages only differ on the surface level, but concepts at a deeper level are the same. Concepts learned in one language can be transferred to a second language. ELLs ideally learn a science concept in their own language without having the burden of simultaneously learning a new language. In this way, the students could transfer their conceptual knowledge to a new language label when learning, for example, English (the L2). Teaching concepts to students in their L1, however, is not always possible due to changing political climates and attitudes toward bilingual education, and educators must therefore grapple with effective ways to teach ELLs science concepts using the English language (Stoddard, Pinal, Latzke, & Canaday, 2002).

Science context. Arguably, all students need to acquire the academic language of science. While researchers (Hudicourt-Barnes, 2003; Lee & Fradd, 1996; Scarcella, 2003) agree that ELL and low-SES students posses a disconnect between their home language and patterns of thinking *and* that of academic language and patterns of thinking in the classroom, they disagree on *how* science instruction should be approached to promote science understanding. On the one hand, researchers believe students' thinking must be scaffolded towards a linear, logical, and structured way of thinking which is, arguably, reflective of real scientific academic language (Lee & Fradd, 1996). On the other hand, researchers believe the creative and flexible thinking that diverse students possess parallels real world scientific thought processes and discourse (Hudicourt-Barnes, 2003). This argument is based on the idea that real-world scientific thinking is not linear, but rather creative and sometimes spontaneous.

Perhaps the best reconciliation between dueling theories in science education has been the idea of using *cultural congruence* or *instructional congruence* (Lee, 2005) in the classroom. In this framework, teachers simultaneously acknowledge, value, and utilize students' cultural and linguistic background (i.e., their everyday language and thinking) while using instructional practices (i.e., explicit teaching of academic vocabulary; modeling of reading non-fiction texts; implementing science writing structures) to scaffold students towards academic language *and* thinking in science. Theoretically, scaffolding with cultural and linguistic sensitivity would allow students to understand science concepts. The ideas parallel theorists (e.g., Cummins, 1981; Halldén, 1999) and researchers (e.g., Fang, 2004, 2006) who delineate differences between academic and everyday language and the critical role academic language plays in conceptual understanding of a content area such as science. The ideas also complement Scarcellas' (2003) more complex sociocultural perspective attempting to bridge the similarities between everyday and academic English in order to scaffold students toward acquiring academic English.

The Role of Writing in Understanding

One potentially potent tool for connecting language, thought, and understanding is the use of discourse, and specifically written discourse, in the science classroom. The following explains the theory and research behind the idea.

Discourse. Vygotsky (1978) believed that learning occurs through linguistic discourse (i.e., speaking and writing). Researchers, in fact, note that when people write about what they have learned, they retain 70% of the content, and when they talk about what they have learned after writing, they retain 90% of the content (Daniels, Zemelman, & Steineke, 2007). Not surprisingly, Rivard and Straw (2004) examined the use of writing and oral discourse in science instruction, concluding that both dimensions of language are critical to learning but that writing was superior to speaking in terms of effectiveness for learning (Rivard & Straw, 2004). The finding aligns with the theories that learning science involves learning a new type of discourse (Lemke, 1990; Sutton, 1996).

Writing as process. Calkins (1994) noted that writing is a process of making meaning in our lives: a tool for constructing understanding as well as for giving students a purpose in learning and thinking. In the context of science education, Yore (2003) noted, "Writing is…a learning tool (technology) that involves students in far more than mere demonstration of knowledge. Rather, the act of writing in science is seen as a process of constructing understanding and building knowledge: the minds-on complement to hands-on inquiries" (p. 712). Keys, Hand, Prain, and Collins (1999) further explained writing in science as a "…process of negotiating meaning in order to construct, refine, alter, and reconstruct science conceptions (p. 1066). In this sense, writing is a tool for creating, and therefore facilitating, understanding.

Writing as representation. Some theorists (Halliday & Martin, 1993) and researchers (Ruiz-Primo, Li, Ayala, & Shavelson, 2004) in science education, on the other hand, have approached writing as a more "static" representation of the academic language and conceptual understanding a student possesses at any given point in time. In this sense, writing is seen less as a process and more as a static representation. The idea is not to discount that writing is inherently an active process, but it allows the researcher to view a written entry as an engraved representation of an individuals' language and understanding at a given moment in time. In this way, researchers can analyze the characteristics of science writing (Halliday & Martin, 1993) and quantify science writing in meaningful ways (Ruiz-Primo et al., 2004). Writing is nonetheless agreed to be a tool for *developing* scientific literacy, which is arguably comprised of academic language and conceptual understanding.

Writing to learn debate. The writing to learn movement, spurred by Langer and Applebee's (1987) work, emphasized writing as a tool to construct knowledge. Langer and Applebee (1987) argued that writing on a topic allows the writer to clarify knowledge, organize ideas, and reflect on learning experience. As has been noted, science education researchers such as Yore (2003) and Keys (2000) embrace the idea of writing to learn, and other science researchers such as Halliday and Martin (1993) view writing in science as specific representations of academic writing acquired by students by being immersed in the discourse of science.

Debate exists regarding which genres (e.g., conventional vs. creative/personal genres) writing should be promoted in school science classrooms. Modernists (Berkenkotter & Huckin, 1995; Halliday & Martin, 1993), on the one hand, claim that students should learn conventional forms (i.e., expository genres such as lab reports or graphic organizers) of scientific discourse to empower them to compete in the mainstream scientific discourse community. Postmodernists (e.g., Prain & Hand, 1996), on the other hand, argue how students should be allowed to write genres that allow more creativity and personal construction and reflection of scientific concepts (i.e., narratives genres including creative writing activities and reflections).

Arguably, a combination of genre pedagogies (i.e., conventional and creative) could prove beneficial for students to acquire both understanding and the language of science as researchers have noted (e.g., Keys, et al., 1999; Akkus, Gunel, & Hand, 2007). Science notebooks, in particular, can become places in which students both record expository processes of science investigations as well as more creative reflections of their learning. Pedagogy allowing for conventional and creative forms of science writing aligns with theorists who acknowledge the complex relationship between everyday and academic language (Scarcella, 2003). The ways in which writing in science has been used in science education research will be explored in depth in the literature review.

Research Overview

Implementing science inquiry teaching methods while integrating science literacy seems to increase ELLs and low-SES students' conceptual understanding and academic language (Lee et al., 2009). The National Science Education Standards (National Research Council, 1996) advocates using science inquiry teaching methods which engage students in activities that allow them to question, investigate, discuss, and share findings about science while the teacher facilitates in the learning process (National Research Council, 1996). Literacy integrated activities in science involve teachers creating opportunities for students to read, write, and speak about science learning in the classroom. Writing in science can take the form of recording questions, ideas, plans for investigation, and findings in science notebooks (Butler & Nesbit, 2008).

Researchers who consider the role of science notebooks on students' language development and impact on conceptual understanding, however, are few (Ruiz-Primo, Li, Ayala, & Shavelson, 2004; Ruiz-Primo, Tsai, & Schneider, 2010). Moreover, only one group of researchers to date have analyzed the role writing plays on the language and conceptual development of ELL and low-SES students in the context of science (Lee, Mahotiere, Penfield, & Maerten-Rivera, 2009). To date, no researcher has specifically examined science notebook entries used in the context of science inquiry from ELL and low-SES students and their role in students' academic language development and conceptual understanding in addition to examining the relationship between language and concept as reflected in their writing. Moreover, only two studies on writing interventions in science with ELL and low-SES student samples consider gender as a variable (Lee, Deaktor, Hart, Cuevas, & Enders, 2005; Lynch, Kuipers, Pyke, & Scesze, 2005). Last, with respect to ELL learners, prominent researchers in science and ELL education have noted the need to further examine the differences in language and concept achievement as reflected in science writing between classifications of ELL learners (i.e., ELLs, former ELLs, and non-ELLs) (Lee, Mahotiere, Salinas, & Maerten-Rivera, 2009).

Purpose of the Study

This quantitative study derived from a randomized, longitudinal, field-based NSF funded research project (NSF Award No. DRL - 0822343) that targeted ELL and non-ELL diverse students from low-SES backgrounds in a large urban school district in Southeast Texas. Researchers from the larger study used science inquiry and literacy integration to promote science and academic language achievement for ELL and low-SES students.

The present study's purpose was to explore how science notebook writing impacts ELL and low-SES students' academic language development and conceptual understanding and how language and concepts are related. Trained raters used a scoring rubric to analyze 5th grade school students' science notebook entries across three individual time points (beginning/middle/end of science notebook kept over the course of one academic year) and across student language status (ELL, former ELL, non-ELL) and gender groups in order to: (a) investigate the impact of science notebook writing on students' academic language development over time, across student language status and gender groups (b) compare students' conceptual understanding across student language status and gender groups, and (c) explore the extent of relationship between students' academic language and conceptual understanding scores.

Research Questions

The following three questions guided my study:

1. Over the course of 1 year, did ELLs, former ELLs, and non-ELL/low-SES 5th grade students make significant gains in academic language, and to what extent does the level of academic language across student language status (ELL, former ELL, non-ELL) and gender groups differ?
2. To what extent does the level of conceptual understanding across student language status (ELL, former ELL, non-ELL) and gender groups for science notebook entries differ?
3. To what extent are academic language and conceptual understanding related as reflected in science notebook entry scores, and how do the relationships compare between student language status (ELL, former ELL, non-ELL) and gender groups?

Significance of the Study

Exploring the impact of writing in science on academic language and conceptual understanding will make a significant contribution to the existing research in science education and will provide insight into the development of academic language and conceptual understanding as reflected in student science notebooks. The study will add much needed research in the field of science education on how academic language and conceptual understanding can be fostered through science notebook writing amongst ELL and low-SES students, who are most at risk academically.

The ultimate goal of education is to acquire knowledge and, in doing so, to develop and apply understandings of the world in society. The United States needs scientifically literate citizens who are capable of understanding and applying knowledge to societal decisions involving science (National Research Council, 1996). If the population is becoming increasingly diverse and our educational system is not working in terms of aiding this population of students to develop scientific literacy, however, then the nation is in danger of falling behind in terms of scientific advancement on a global scale.

For ELL and low-SES students, language and concept are theoretically closely related and therefore critical to examine. As Vygotsky (1978) argued, without language, it is not possible to learn. Thus, if ELL and low-SES students do not understand the academic words being used in the classroom, they will fall behind in conceptual learning and understanding (Cummins, 1981).

In the context of science, the fusion of language and concepts is critical in an era of high-stakes testing, where academic language is the key to success for these students. As researchers pioneering the field of science learning with ELLs and low-SES students have observed, the level of the students' language proficiency parallels their ability to build upon science understanding as well as to demonstrate it (Kieffer et al., 2009; Lee

& Lyukx, 2005). For this reason, there is a strong call to integrate science and literacy instruction for students on the part of researchers (Fang, 2006; Janzen, 2008, Lee, 2005).

Educators and researchers, however, must be able to accurately and efficiently assess student progress in order to scaffold students to the next level of understanding. In the realm of instruction including writing in the science classroom, only one group of researches have attempted to create a robust assessment tool to measure students' science notebook writing (Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010), and the researchers do not specifically mention ELL or low-SES students in their sample. Combining the use of a robust science notebook writing rubric with samples of ELL and low-SES student science notebook entries will provide much needed and critical insight into the academic language development and conceptual understanding of these students in the context of science inquiry instruction.

Limitations

The present study includes limitations which readers should be aware of when interpreting results. First, the sample included fifth grade students from low-income families, some of which were ELLs, in a single community. The sample may therefore not represent the population of all fifth grade, low-SES and/or ELL students and results are not generalizable. Generalizability, however, may be inferred to students with similar characteristics to those in this study. Second, though the larger study was quasi-experimental, science notebook writing samples from the control classrooms were not collected for the present study because the control classrooms were not expected to implement science notebook writing in science. It is not possible, therefore, to decipher whether student results are due to the intervention instruction. The analysis, however, can provide critical insight into the role science notebook writing played on students' academic language development and conceptual understanding within a science inquiry intervention.

Delimitations

Even though the science curriculum in the present study covered four science domains, the nature and scope of the present study delimits to an examination of three science domains. First, the study is exploratory – no previous study has attempted to explore ELL and low-SES's academic language development and conceptual understanding from science notebook samples. Second, science notebook ratings are laborious. Therefore, rating three entire units (one from the beginning of the year, one from the middle of the year, and one from the end of the year) was sufficient for calibrating the instrument and obtaining data needed to track academic language growth over time.

Assumptions

An assumption of the present study is that teachers knew and implemented the science inquiry curriculum from the MSSELL grant well. Teachers implementing the science intervention received biweekly professional development. In addition, teachers were observed and evaluated on the fidelity of classroom implementation. Researchers took fidelity observations using the Science Teacher Observation Record (STOR) at the beginning, beginning/middle, middle, and end of the school year. The observers rated the teachers on a scale of 1-4 on their (a) knowledge with lesson content, (b) material usage and teacher preparation, (c) student involvement, (d) academic language scaffolding, (e) affective and cognitive feedback, (f) writing feedback, and (g) pacing. The assumption that the curriculum was implemented with fidelity is therefore reasonable.

Organization of the Study

Chapter I of my study includes the definition of terms, a statement of the problem, the theoretical framework, the purpose of the study, the research questions, the significance of the study, the limitations/delimitations, and the assumptions.

Chapter II of my study includes an introduction, writing and conceptual understanding in science, writing and academic language in science, the relationship between concept and language in science writing, a discussion, and a conclusion.

Chapter III of my study includes an introduction, sample, setting, research design, instrumentation, intervention procedure, data collection, data analysis, and a summary.

Chapter IV of my study reports the data analysis and summary.

Chapter V of my study presents a discussion of findings, limitations, and conclusions and recommendations.

CHAPTER II

LITERATURE REVIEW

The purpose of this literature review was to examine literature on (a) the impact of writing in science on students' conceptual understanding of science, (b) the impact of writing in science on students' academic language, and (c) the relationship between concept and language as reflected in students' science writing.

In order to locate pertinent research on the topic of the effect of writing in science on students' language development and conceptual understanding, 14 peer-reviewed educational journals publishing studies in science, instruction, and bilingualism (See Appendix A for a list of the journals) were searched using the same combinations of key terms and connectors (i.e., writing, science, science notebook, literacy, science literacy, English Language Learner, science inquiry, concept*, understand*, language develop*). Initial search parameters included peer-reviewed studies published within five years (2006-2011), having to do with writing in the science classroom (preferably inquiry based learning but not limited to that setting), and with students in the middle grades of low-SES or diverse language (i.e., ELL) backgrounds in the U.S. This search yielded 7 empirical studies.

From purling the 7 retrieved article references, it was clear that the search needed to be adjusted to include seminal studies that were published more than five years ago as well as studies that included non-ELL students at any grade level in any country. Given that the *National Education Standards* were published in 1996, the search parameters were adjusted studies published within 15 years (1996-2011). Studies included elementary through high–school students of any background (i.e., homogenous or diverse; low SES or middle/high SES) and conducted in any country but published in English. This search yielded a total of 12 more empirical studies on the topic of science and literacy integration published within the 14 peer-reviewed education journals, for a total of 19 studies.

In total, the search yielded 5 empirical studies on the role of science inquiry interventions that included some element of science literacy integration and a focus on student achievement in science and literacy (vs. teacher change or student science inquiry ability) and 14 empirical studies specifically on the topic of writing in the science classroom (N total = 19). The studies are synthesized below.

Analysis of the 19 studies yielded three main categories: (a) writing and conceptual understanding; (b) writing and academic language; and (c) relationship between language and concept in science. It should be noted that some studies intersect the three main categories. In these cases the studies are discussed under each category. The studies are further organized according to whether the sample included ELL students or not and the writing pedagogy used in the study; that is, whether teachers had students use conventional science writing-to-learn strategies (e.g., expository writing or lab reports), creative/personal science writing-to-learn strategies (i.e., narratives or other forms of creative writing and/or reflections), or a mixture of both strategies. Literature review matrices are in Appendix B.

Writing and Conceptual Understanding

The largest sample of studies found included researchers exploring the effect of writing on students' conceptual understanding in science, and, all included writing pedagogies (i.e., writing-to-learn strategies) embedded within science inquiry units (N = 14). As has been defined, conceptual understanding refers to the ability to be able to use science knowledge (i.e., facts, concepts, principles, laws, theories, and models) (National Research Council, 1996).

Some researchers have noted that standardized measurements, which are inherently less aligned with the classroom curriculum, tend to be less accurate measures of students' conceptual understanding than performance-based assessments that are more closely to the curriculum (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). However, standardized assessments, if used in the studies reviewed to measure students' science knowledge and understanding, are acknowledged in this review as reflections of conceptual understanding.

It is my shared belief that performance-based assessments (such as writing entries in science notebooks) are ideal measurements for students' science conceptual understanding (Ruiz-Primo et al., 2002). The reality, however, is that (a) standardized assessments are viewed by policy makers as reflections of students' understanding (knowledge and concept) of a subject area and (b) standardized tests do require students' application of knowledge in answering questions, and therefore the use of knowledge, which is part of the definition of conceptual understanding (National Research Council,

1996) and (c) quantitative studies seeking to find the impact of writing on science conceptual development often rely on standardized assessments as is discussed below.

Studies with Non-ELL Samples

The majority of the studies considering the effect of writing on students' conceptual understanding (N = 9) included sample populations of non-ELL students from middle class, homogenous backgrounds. The samples, however, were from diverse contexts including two studies in the U.S.A. (Gunel, Hand, & McDermott, 2009; Keys et al., 1999), two in Northern Italy (Mason, 2001; Mason & Boscollo, 2000), one in Southern Wales (Patterson, 2001), one in Turkey (Hand, Gunel, & Ulu, 2009), one in Australia (Ritchie, Tomas, & Tones, 2011), and one in a French-Canadian province (Rivard & Straw, 2000). Researchers in one study did not mention a context (Akkus, Gunel, & Hand, 2007). Within this category, three studies were qualitative, five were quantitative, and one was mixed-methods. Two studies included conventional science writing-to-learn strategies, five studies included creative/personal science writing-to-learn strategies, and two studies included mixed forms. None of the studies included the use of science notebooks. (See Appendix B for a matrix that organizes studies in this category).

Conventional science writing. Researchers using conventional science writing-to-learn strategies (e.g., expository writing; lab reports) reported positive results from scaffolding students' ideas as a way to encourage scientific understanding (Gunel et al., 2009; Hand, Gunel, & Ulu, 2009; Patterson, 2001; Rivard & Straw, 2000). The researchers' scaffolds, however, differed. Hand et al. (2009) and Gunel et al. (2009)

conducted studies that required students to write expository explanation letters to students in younger grades. In a sense, the practice allowed students to scaffold their own scientific language as they had to think about how to clearly convey their message to a less experienced audience. Patterson (2001) used concept maps to help students organize their ideas before writing them down, and Rivard and Straw (2000) used combinations of student "talk" and writing to guide students through a problem-solving task. The studies are described in more detail as follows.

Gunel et al. (2009) noted, "…writing-to-learn activities help students gain conceptual understanding of scientific topics" (p. 364). In their study, the researchers wanted to know whether 20 9th grade and 98 10th grade students in four different classes in an upper/middle class Midwestern U.S.A. class would perform differently on posttests of science understanding if they wrote explanations of the science topics they were learning about (i.e., the circulatory and respiratory system) to different audiences (i.e., to a younger audience, to their teacher, etc.). As a result of their Analysis of Covariance (ANCOVA) and Multiple Analysis of Variance (MANCOVA) analysis of posttest scores on researcher-developed unit tests and instructor-created rubric scores of student writing, the researchers noted that students who wrote to a 3rd and 4th grade audience scored significantly higher on the conceptual questions on posttests than students who wrote to their teachers. The researchers noted that perhaps having to write to a younger audience forced the students to translate their conceptual understanding into simpler language, thus re-enforcing their own understanding of the science concepts.

In Hand et al.'s (2009) study, the researchers also had their students write science explanations to a younger audience, noting positive results in their students understanding at the end of the process. However, the researchers' focused on whether the additional scaffold of having students embed mathematical representations into their writing would help their comprehension of the science topic. The study was a quasi-experimental pre-posttest study with 172 tenth-grade male students in a semi-private boarding school in Turkey. Students were asked to write explanations of their science topic to the 9th grade students in their school. One group used only text to write; one group used text and math; and one group used text and a graph. ANCOVA analysis comparing the mean differences between the groups supported the pattern of advantage of embedding text plus mathematical representations in writing. Scaffolding conventional science writing with math therefore helped the students on the posttest. It should be noted, however, that samples of posttest questions provided in the study showed a heavy emphasis on math. The fact that the writing tasks with math were closely aligned to the posttest items likely had an impact on the scores. This is not a negative point; however, how closely aligned a writing task and its measurement instrument is to the task, can affect the statistical significance of results.

Patterson (2001) qualitative study used concept maps as a way to scaffold students' thinking prior to writing about the science topics they were learning. Patterson's (2001) sample students in Southern Wales (N = 6) in years 2, 3 & 6 learning science. The researcher explicitly taught the students how to use concepts maps before writing. After analyzing the students' science writing, the researcher noted that "…the process of writing can enhance pupils' learning in science", and, furthermore, that providing scaffolds such as concept maps, allowed students to "…demonstrate far greater concept understanding in their writing" (p. 15).

Rivard and Straw's (2000) seminal, quasi-experimental study considered three different scaffolding interventions used in a French language school in a homogenous province made up of low middle class to upper middle class families in which French was used as the first language (note: despite its context the study does not classify itself as bilingual or ELL focused). Forty-three eighth grade students were randomly assigned one of the intervention groups – in one intervention group, the students discussed a problem task in small groups; in the second intervention group, the students wrote individual responses for a problem task; and in the third intervention, the students both discussed the problem task in a small group and then wrote individual explanations. The researchers measured student science learning/ understanding with a multiple-choice test, a short essay question test, and the creation of concept maps. Using rubrics, descriptive analysis, and analysis of covariance, the researchers explored the role that speaking, writing, and a combination of both played in students' learning. They concluded the following: a) Talk is important for sharing, clarifying, and distributing knowledge; b) Analytical writing is a tool for transforming ideas into more coherent and structured knowledge; c) Talk combined with writing enhances retention of science learning over time. So, while speaking and writing both play distinct roles in scaffolding students' understanding, it appears that using both as scaffolds for students' understanding is ideal.

Creative/personal science writing. Researchers (Mason, 2001; Mason & Boscollo, 2000; Ritchie et al., 2011) considering creative/personal science writing have also noted the benefit of allowing students to collaborate by discussing science activities and then writing about them. For example, Mason and Boscollo (2000) and Mason (2001) conducted studies in which the researchers promoted student peer collaboration through discussion about their science learning experiences combined with writing down reflections on their understandings of the content material. Peer collaboration and individual writing served as scaffolds for students' conceptual understanding during science lessons. Other researchers such as Ritchie et al. (2011) highlighted the benefits of introducing creative forms of narrative story writing in science in order to solidify conceptual understanding. The studies are discussed in greater detail as follows.

In their mixed- methods study, Mason and Boscollo (2000) explored whether writing in the science classroom improves the understanding of a new topic. The instruction involved teachers modeling writing, students being asked to record, reflect, and express on experiments, and students using writing to link to new concepts being learned. The study was conducted in Northern Italy with a classroom of 12 fourth grade students from a homogenous, middle class background. Three fourth graders (n = 16 in experimental and n = 20 in control) participated in the study. Control and experimental groups received the same curriculum and instruction on science units for two and a half months; experimental included writing for learning instruction. Quantitative evidence from ANOVA results showed that the experimental group reached higher levels on all the posttest science measures than did control students. Qualitative analysis of the experimental group children's written tests led the researchers to conclude that writing "helped [the student] to better understand the new topic…" (p. 222).

Mason's (2001) qualitative study, though also focused on personal writing reflection, did not include structured writing instruction (i.e., students were not explicitly taught a writing strategy). In the study, 12 fourth-grade students in Northern Italy were given opportunities to discuss and write before, during, and/or after engaging in science activities. Mason (2001) conducted qualitative analysis of the students' samples, noting that the students advanced at different levels of scientific understanding. Thus, it appears that while writing was can be a tool for students to reflect on ideas and experience and therefore refine their understanding, the tool benefits from being explicitly taught and modeled like in Mason and Boscollo's (2000) study.

Ritchie et al.'s (2011) study had 55 sixth-grade students (29 treatment and 27 control) in a "well-resourced suburban Australian public school" (p. 690) creatively write narratives, developing characters and scenarios to explain the science topics they were learning. The narratives required some creative thinking as the students developed characters and a scenario to explain the science concepts. Students in both treatment and control groups took a pre and posttest *BioQuiz* to measure their science literacy, and MANOVA results revealed that students in the treatment class showed improvement on posttest *BioQuiz* scores compared to the treatment class. Posttest narratives, moreover, showed a significant mean difference in mean science content scores. The researchers thus implied a link between the narrative writing task process and students' scientific literacy (i.e., knowledge and interest on the science topic) and conceptual science understanding.

Mixed science writing. A final group of studies in this category attempted to combine the use of conventional science writing and more creative and personal forms of writing by means of a program called the Science Writing Heuristic or SWH. Akkus et al. (2007) explain the SWH as a “bridge between informal, expressive writing modes that foster personally constructed science understandings and more formal, public writing modes that focus on canonical forms of reasoning in science” (p. 1746). The researchers explain how the program provides a template for teachers in which they guide students through several phases which include constructing and testing questions, justifying claims with evidence, reflecting on how ideas have changed, and a writing task for the purpose of negotiating and clarifying meaning while producing a final task. The SWH emphasizes the role of collaborative student work through class discussion and focuses on developing students’ deep understanding of science concepts. Knowing this, the SWH is, in itself a structured lesson plan template and scaffold meant to guide students’ conceptual understanding through the lesson sequence.

In their study, Akkus et al. (2007) conducted a treatment-control pre-posttest study with 592, 7-11th grade students (270 control and 322 treatment; context not given) in which the treatment group received Science Writing Heuristic (SWH) instruction. Pre and posttests were teacher-generated unit specific tests, and ANOVA and ANCOVA results indicated that students in the treatment group scored significantly higher than students in the control group if and only if the teacher provided quality implementation of the SWH approach. Furthermore, achievement gaps between high achieving and low achieving groups disappeared when the SWH was implemented well. Similarly, Keys et al. (1999) explored whether using the SWH over the course of an 8 week science unit engaged students in a learning process that led to conceptual change in their understanding of science. The sample included nineteen eighth grade students from two eighth-grade classes in the Southeastern United States. The researchers collected and analyzed students' writing samples. They concluded that students engaged in meta-cognitive thinking – "writers reflected on the sources of their knowledge, the degree of certainty of their knowledge, and how their knowledge had changed over time" (p. 1081) - to understanding and meaning of the scientific data they encountered. Thus, it appears that the SWH is a useful tool for allowing students a space in which to write reflections and conventional science writing forms and is effective for developing students' conceptual understanding.

Up to this point, the studies discussed have included samples from homogenous, mostly middle class students. The following will discuss studies also concerned with students' science conceptual development but that include samples with ELL and/or low-SES students.

Studies with ELL Samples

Studies with ELL samples (N = 5) related to students' conceptual understanding in science were all quantitative and fell within larger inquiry and science-integration intervention studies. Science achievement scores in these studies were viewed as indicators of the impact of literacy integration, including writing, on students' understanding of science. As will be discussed below, results were overall positive for ELL students when the inquiry and literacy science integration programs include culturally sensitive instruction. Three studies (Fradd, Lee, Sutman, & Saxton, 2001; Lee, Deaktor, Hart, Cuevas, & Enders 2005; Lee, Maerten-Rivera, Penfield, LeRoy, & Secada's, 2008) used mixed forms of conventional and creative/personal writing-to-learn strategies, and two studies (Amaral, Garrison, & Klentschy, 2002; Lynch et al., 2005) used of science notebooks in the classroom (it appears the notebook entries included conventional writing-to-learn entries). The studies are discussed below. (See Appendix B for a matrix that organizes studies in this category).

Mixed science writing. Lee and colleagues' work (e.g., Fradd et al., 2001; Lee et al., 2005; Lee et al., 2008) has been seminal in the area of science and literacy instruction with ELL students. The following three studies include Lee as an author and are therefore grouped together. As has been noted, the studies are embedded within larger science-inquiry interventions. All of the studies mention students writing expository and/or narrative paragraphs about their science topics and experiences as well as creative responses to prompts.

Fradd et al. (2001) synthesized the results of two projects that developed science materials aimed and promoting science inquiry and cultural congruence for ELLs and developing science literacy in the United States: the *Promise Project* and *Science for All.* In the interventions, students were instructed to, "Record what you did so others can learn. Consider different ways to express your information" (p. 491). The researchers do not provide any more information about student writing, so it can only be inferred that students were given freedom to think of different forms of writing to present their results (conventional and/or creative). Data analysis for 4th grade students participating in the project included pre and posttest unit science tests. The *Promise Project* included a treatment and control group (though the summary section is not clear about how the groups were assigned), and researchers found students using inquiry units to outperform students not using the units on post unit science test scores as reported by descriptive t-test analysis. *Science for All* included pre and posttest scores (no intervention and control comparison), and all students participating in the program showed statistically significant gains in post unit science test scores according to descriptive t-test analysis.

In a later publication Lee et al. (2005) reported the impact of an instructional intervention involving science inquiry, English language and literacy integration, and home language and culture on ELL students' science achievement. Students were asked to write expository and narrative paragraphs about science processes and experiences in addition to responding to science writing prompts. Again, the researchers do not elaborate more on the students' writing except to explain the pre and post test writing prompt as a combination of expository and creative writing in which the students were to pretend they are drops of water and then explain the water cycle (p. 867). Students in the study included third and fourth grade students in six participating public elementary schools in the United States (n = 1,500), and part of the intervention including having students write expository paragraphs or narrative stories to describe the science processes under investigation. District developed science unit tests served as pre and posttests for student science achievement, and the researchers found that students demonstrated statistically significant gains and large effect magnitudes (Cohen's *d*) on all measures of science achievement at the end of the school year. While the growth rates for ELL, low-SES and non-ELL, and high SES student were the same, ELL students and low SES students performed significantly lower than mainstream, high SES students on pre and posttests. Thus, ELL and non-ELL growth rates were equal, but the achievement gain did not close between the groups.

Lee et al. (2008) found similar results in their study on the impact of the first year intervention of professional development on the science achievement of ELL students. Again, students were asked to write expository paragraphs to describe science processes, explanations, or conclusions in addition to responding to science unit prompts. Little else is said about the students' writing. The students were third-graders from United States public urban schools in a low-SES setting. The study included a treatment (n = 1, 134 students) and control group (n = 959 students) that were assigned to the conditions based on specific criteria. Pre and posttests developed by the researchers to measure students' understanding of science concepts were administered only to the treatment students and only statewide math achievement tests were compared between treatment and control groups. It is not possible to conclude whether the results of the science intervention increased science achievement for treatment over control groups since science measures between the groups were not compared. However, amongst students receiving the treatment, HLM analysis found statistically significant science achievement gains from pretest to posttest. Moreover, the researchers did not find statistically significant differences in achievement gains between ELL students and students who had exited from ELL status or had never been ELL (i.e., all student groups made similar gains which is a positive finding); though, again, the achievement gap did not necessarily close. These findings, like the ones from Lee et al.'s (2005) previous study, confirm the importance of early and intensive intervention for ELL students in order to close or all together avoid achievement gaps.

Science notebook writing. In Amaral et al.'s (2002) study, students were expected to use science notebooks to "… collect, record, analyze, and report data for each of the inquiry units" (p. 224). The notebooks appear to consist of conventional writing entries. In addition, the researchers explain that the purpose of science notebooks was for students to develop cognitive knowledge and English language skills. The study was seminal in that it was one of the first longitudinal, science inquiry studies with ELLs from a low SES population. Specifically, the study was conducted in public schools in Southern California with a final sample of 615 fourth grade students and 635 sixth grade students that participated in the program for all four years. Students in the program received kit and inquiry based science instruction that included the use of science notebooks, and the teachers received professional development. The study measured students' science learning with the Stanford Achievement Test, ninth edition, Form T (4th and 6th grade). At the end of the study all students, including ELLs, had higher achievement scores in science the longer they were exposed to the program.

In Lynch et al.'s (2005) quasi-experimental study, students in the intervention group received a structured curriculum unit for 6-10 weeks with guided inquiry lessons and the use of science notebooks. The science notebooks were used to "analyze results, and use evidence-based arguments in large and small groups to support their claims" (p. 921). Again, it appears the notebooks included conventional writing entries. Lynch et al. (2005) looked at a group of 1,500 eight grade students across five public middle schools in the United States servicing low-SES students. A team of science content experts, educators, teachers, and assessment specialists created pre and posttests to measure student science achievement. Based on their ANCOVA analysis results, the researchers noted that students in the treatment group, overall, showed statistically significant posttest results for achievement in science and outperformed the comparison group. ELL students, however, did not outperform the comparison group. The researchers attribute the ELLs' lack of performance to the high literacy demands of their program and/or the assessment which could have "failed to capture the learning gains of these students" (p. 942). Certainly, deeper analysis of the outcomes could be conducted to understand where ELLs struggled. At the same time, as August, Branum-Martin, Cardenas-Hagan, and Francis (2009) note, Lynch et al. (2005) did not make any mention of linguistic or cultural alterations to the curriculum, unlike the other studies in this category. This could be a noteworthy point in terms of effective instruction for ELL students. Perhaps science inquiry alone is not effective for ELLs. However, Lynch et al.'s (2005) pointed out the uncertainty of whether their assessment captured the learning gains due to its high level of English use. Researchers interested in the performance of ELLs on standardized exams acknowledge the critical role and possible barrier that language plays (Kieffer et al., 2009). Further examination of researcher-developed assessment instruments in this study is warranted.

Writing and Academic Language

Studies that fall into this category tend to use diverse measurements (i.e., qualitative linguistic analysis or researcher-created rubrics) to measure students' academic language. For this review, the studies (N = 5) were included as long as the researchers sought to answer the following question: How does writing in science impact students' academic language?

Studies with Non-ELL Samples

Studies with non-ELL samples are very few (N = 2). Only one qualitative (Patterson, 2001) and one quantitative (Ruiz-Primo et al., 2004) study fall within this category, and one (Patterson, 2001) has been discussed under the category of writing and conceptual understanding. Perhaps what most stands out in this category is that the one study that used a conventional science writing-to-learn strategy had positive outcomes in terms of students' academic language (Patterson, 2001), while the study that did not use a specific writing-to-learn strategy (i.e., the study mentioned the teachers were expected to science notebooks in the classroom, but the researchers to do not mention how the writing was taught) did not have positive outcome in students academic language (Ruiz-Primo et al., 2004). The studies are discussed below. (See Appendix B for a matrix that organizes studies in this category).

Conventional and specific writing. As has been previously noted, Patterson (2001) taught a sample of students in Southern Wales (N = 6) in years 2, 3 & 6 to organize their thinking using concept maps before writing. In addition to exploring whether writing in this manner would impact scientific understanding, Patterson (2001) explored whether the writing strategy would help students create more coherent writing. Teachers in the intervention qualitatively analyzed student writing after using context maps to help them plan. The researchers reported that the use of context maps increased the quantity of writing as well as the use of connective words and number of explanations. The findings were interpreted as indicators of higher quality writing, under the assumptions that more words meant more fluency and that the presence of connective words and explanations reflected the more effective and sophisticated use of language (vs. simple descriptions).

Science notebook writing. While Patterson's (2001) qualitative findings were promising, Ruiz-Primo et al.'s (2004) quantitative findings were less promising. The researchers analyzed a sample of 72 science notebooks randomly selected from six fifth-grade classrooms. From the researchers' description of the writing genres they identified in the notebooks, the notebooks seem to have included mostly conventional writing entries (i.e., expository descriptions of processes; hypothesis; lab reports) with a few creative writing entries (i.e., narratives and reflections).

The classrooms were located in a school district in the Bay Area in California in the United States and had used a science inquiry curriculum as part of a larger study. The researchers described the development of a rubric to rate the science notebooks and rated the notebooks on several criteria (understanding, opportunity to learn), including *quality of communication.* In a pilot study, the *quality of communication* criteria referred to the completeness, clarity, and organization of the writing in general. The researchers, however, noted that the resulting student scores from the pilot study had little variation, thus concluding that, “The criteria did not accurately discriminate the quality of communication across students” (p. 1483). As a result, they changed the scoring criteria to consider how well students’ writing entries aligned with the genres of scientific communication based on Lemke’s work (1990) on scientific genres (i.e., description, comparisons, definitions, lab reports, etc.).

The researchers analyzed the kind of genre the writing entry represented and how well the students’ language aligned with linguistic characteristics of the genre. For example, if an entry was classified as a definition, the researchers used a rubric that rated the completeness of the definition including specific aspect of language characteristic of that genre such as the use of technical terms and verbs in the present tense. T-test analysis of overall *quality of communication* for students’ writing led the researchers to conclude that his criteria did not improve over time, “…due, in part to the fact that no teacher feedback was found in any of the students’ notebooks. Therefore, there was no effort to close the gap between the student performance at the time that the notebooks entry was produced and the desired performance” (pp. 1500-1501).

Ruiz-Primo et al.’s (2004) findings reflect Akkus et al.’s (2007) study in which the researchers found positive results in student conceptual understanding if and only if the teachers implemented writing instruction well; otherwise, the results were not statistically significant like in Ruiz-Primo et al.’s study (2004). It appears, then, that writing in science in itself may not be enough to positively impact students’ academic language conceptual understanding; rather, teacher involvement and instructional implementation (i.e., teaching the writing-to-learn strategy) plays an important role in the use of writing as a tool in the classroom for learning.

Studies with ELL Samples

Studies examining the impact of writing in science on ELL students' academic language are very few in number (N = 3). The studies are, like the ELL studies previously discussed on science conceptual understanding, embedded within larger science inquiry studies and all the studies use quantitative analysis. The researchers in the studies concur that the longer ELLs were exposed to writing in science, the more gains they made in their science and/or English language development. Students in the studies were reported to either a mix of expository paragraphs and creative writing prompts (Lee, Deaktor, Hart, Cuevas, & Enders, 2005; Lee, Mahotiere, Salinas, Penfield, & Maerten-Rivera, 2009) or the use of science notebooks in the classroom, seeming to consist of only conventional writing entries (Amaral et al., 2002). (See Appendix B for a matrix that organizes studies in this category).

Science notebook writing. As has been previously noted, Amaral, et al.'s (2002) longitudinal study implemented science inquiry and literacy integrated curriculum in a school district made up of ELL and Low-SES students. The intervention specifically noted the structured and expected use of science notebooks in the classroom, which appears to have included conventional forms of science writing. In addition to measuring students' science learning, as has already been noted, the researchers also measured students' writing (i.e., language) development. The researchers used a district writing proficiency test administered each winter and spring during the four-year intervention.

Classroom teachers collected the writing tests and trained evaluators used a four-point rubric covering content and the conventions of writing to score the writing tests. Descriptive statistics on the number of students who passed the district writing benchmark showed that the longer the students were in the program, the higher their writing scores. For example, students in Grade 4 at the beginning of the intervention had a 57.6% passing rate, but by grade 6, they showed an 86.8% passing rate. Unfortunately, information about the writing rubric, process, and inter-rater reliability were not provided.

Mixed science writing. Lee et al.'s (2005) study, as has also been previously discussed, reported the impact of the first year of a science inquiry and literacy instructional intervention with third and fourth grade students in the United States (n = 1, 500). As part of the intervention, students wrote expository paragraphs or narrative stories to describe the science processes under investigation. For writing measurement, students completed a pre and posttest writing prompt. Researchers rated the writing samples using a researcher-created writing rubric, reported to have 90% inter-rater reliability. Based on mean comparisons between pre and posttests (with t-tests) and effect magnitudes (Cohen's *d*), the researchers concluded that students demonstrated statistically significant gains and large effect magnitudes in writing achievement measures at the end of the school year.

In Lee et al.'s (2009) study, the researchers focused on analysis of the writing samples from all three years of the intervention described in their 2005 study. As they explain in their previous study, the writing samples included a pre and posttest response to a writing prompt in which the students were asked to pretend they were drops of water and explain the water cycle (i.e., a mix of expository and creative writing). The researchers used a sample of all third-grade students from six treatment schools during the first 3 years of their larger study (2004-2007), reporting 683 students in year 1; 661 students in year 2; and 676 students in year 3. Teachers collected the writing prompts and the researchers used two scoring rubrics that they developed to assess *form* (i.e., conventions, organization, and style/voice) and *content* (i.e., knowledge and understanding of the water cycle presented in the third grade curriculum). Both rubrics consisted of a five-scale system. Researchers report a 90% inter-rater reliability. As a result of their HLM and HGLM analysis, the researchers concluded that students made significant achievement gains each year for form and content, and that the gains were incrementally larger for writing form. Again, ELL students made achievement gains comparable to ELL exited and non-ELL students but ELL students had lower form and content scores than students who had exited from ELL programs or had never been classified ELL. This study stands out because it compared student language gains across student classifications (unlike the other two in this category) – a critical task if researchers are to further unravel the role of language and thinking for ELL students.

Relationship between Academic Language and Conceptual Understanding

Studies in which researchers specifically consider at the relationship between academic language and conceptual understanding are also small in number (N = 5). The studies in this category sought to either qualitatively analyze how linguistic features characterize student understanding (Keys et al., 1999; Keys, 2000) or to quantify correlations between language scores and conceptual scores (Gunel et al., 2009; Ruiz-Primo et al., 2004; 2010). Notably, all of the researchers agreed that some relationship exists between language and concept, but none of the researchers used samples from populations with ELL students to analyze this relationship. Researchers report conventional writing-to-learn strategies in two studies (Gunel et al., 2009; Keys et al., 1999), a mix of conventional and creative writing-to-learn strategies in one study (Keys, 2000), and the use of science notebook writing in two studies (Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010). (See Appendix B for a matrix that organizes studies in this category).

Conventional Writing

In their first study considering writing in science, Keys et al. (1999) found that linguistic structures such as inferences and linguistic patterns such as expansion and generation of ideas were related to science understanding. Students worked with a partner to compose written reports about their observations while taking part in a summer science camp inquiry project; the students did not receive explicit writing instruction. The students came from 34 middle school students (33 African American; 1 Latino) from five urban schools (4 low SES; 1 middle class) in the Southeastern United States. The researchers conducted content analysis of student reports using functional grammar analysis (i.e., how language is used to achieve its purpose) to classify the type of information the students generated and found that few students produced the linguistic structures related to science understanding and commented on the importance of explicitly teaching students scientific discourse. The researchers' conclusions echo those of Akkus et al. (2007) and Ruiz-Primo et al. (2004): writing instruction is key for students to learn how to construct language in an effective manner on paper to communicate scientific ideas.

Gunel et al.'s (2009) quasi-experimental pre-post study, previously discussed, included a writing-to-learn strategy implemented with high school students during science instruction in which the students wrote expository pieces to explain concepts to different audiences. In addition to positive findings on the impact of writing on student conceptual understanding, the researchers also conducted a regression analysis with treatment students' writing scores and posttest science score. The researchers found writing assignment scores to be significant predictors for student performance on science posttest measures. Writing and concept are therefore possibly related. The study noted explicit instruction and student feedback on the part of the teacher.

Mixed Science Writing

Keys (2000) explored the use of an explicit writing instructional strategy (the Science Writing Heuristic, or SWH, previously described) from a science classroom lab activity, noting that certain uses of language (i.e., generation of explanations; reflections) lead to science learning while other language uses (i.e., those that are purely descriptive) do not. In the study, 16 eighth graders were chosen as subjects from a rural middle school in the Southeastern United States. Keys (2000) qualitatively analyzed the students' scientific reports by classifying the students' writing into thematic interpretations. Her findings included the following: (a) some students generated new knowledge and explanations form writing and some did not (b) some students reflected while writing (c) students who focused on problem solving generated higher levels of scientific thinking and learned science from writing. It is not clear why only some students were able to use writing effectively for the formation of new knowledge, but her last finding provides a spin on how language and understanding may be related to thinking. If problem solving led to the generation of higher quality thinking and writing, then context matters in terms of how language is learned, used, and applied.

Science Notebook Writing

Ruiz-Primo et al.'s (2004) study also found relationships between language and concept as reflected in writing notebook scores. As has been discussed in the previous section, the researchers created a rubric to rate a sample of 72 randomly selected science notebooks from six fifth-grade classrooms that took part of a larger study in science inquiry and literacy integration. The rubric rated *quality of communication* (i.e., how well student writing entries aligned with genres of scientific communication; described under this literature review's section, *Writing and Language Development*) and *understanding*. According to Ruiz-Primo et al. (2004), understanding refers to either conceptual or procedural understanding. As the researchers explained, if an entry focused on "defining, exemplifying, relating, comparing, or contrasting unit-based concepts" (p. 1484), they considered it to reflect conceptual understanding. If an entry focused on "reporting procedures carried out during an activity/experiment, reporting observations/ results/ outcomes, interpreting results, or concluding" (p. 1484), they considered it to reflect procedural understanding. Each type of *understanding* had a separate four-point rubric, like the rubric for *quality of communication*. A Pearson *r* correlation analysis between the scores for *understanding* (i.e., concept) and *quality of communication* (i.e., language) indicated positive correlations. The researchers furthermore found positive correlations between composite notebook scores and other student performance indicators (i.e., close, proximal, and distal assessments). Language and concept are arguably related; and writing, as a whole, is arguably correlated with learning. It should be noted, however, that Ruiz-Primo et al. (2004) extensively discussed that though the correlations were present, the notebook scores were not necessarily high. The researchers attributed this fact to the lack of evidence of effective use of the notebooks in the classroom (i.e., little teacher feedback).

In a more recent study Ruiz-Primo et al. (2010) also found relationships between language and concept in student writing. In their study the researchers sought to explore the link between the quality of student writing and students' learning achievement as reflected in science notebook explanations of lab experiments. The researchers specifically explored the quality of student explanations and learning achievement within a science unit. To do this, nine student notebooks from within eight middle school classrooms in the United States were randomly selected and rated by a researcher-created rubric. Descriptive statistics showed that the level of students understanding was consistent with the quality of students' explanations; and correlational analysis of the quality of student's explanations to other performance indicators (i.e., post-test assessments) showed positive to moderate correlations. The researchers therefore concluded that engaging students in writing has a positive impact of student learning of content; though, the researchers noted that student levels of understanding and quality of writing are affected by the teachers' implementation of science and literacy instruction.

Discussion

The following discusses the studies within each of the categories from the findings (i.e., *Writing and Conceptual Understanding*; *Writing and Academic Language*; and *Relationship Between Language and Conceptual Development*). The discussion highlights differences between the studies' context, design, writing interventions, and methods within each category and provides suggestions for future research in the area of writing and science instruction. In addition, separate sections discuss studies that differentiated student achievement according to student language classification and gender and studies that consider using rubrics to rate science writing.

Writing and Conceptual Understanding

Studies focused on the impact of writing on students' science conceptual understanding were by far the largest in number (N = 14). Researchers agreed that writing had a positive impact on students' conceptual understanding of science, if strong scaffolds were in place in the instruction.

The study contexts, designs, interventions, and measurements differed; and the variations are worth noting for drawing conclusions and considering research directions for the future. The following discusses these differences by comparing studies with non-ELL samples to studies with ELL samples and providing suggestions for future studies.

Non-ELL and ELL study contexts and designs compared. Studies with non-ELL samples (N = 9) outnumbered studies with ELL samples (N = 5). The research contexts and designs were strikingly different between the two categories as is discussed below.

Contexts. Studies with non-ELL samples were conducted in diverse, international contexts with mid to high-SES students while studies with ELL samples were all conducted in the United States with low-SES students. Also, with the exception of Mason (2001) and Mason and Boscolo's (2000) studies with 4th grade students, all of the studies with non-ELL samples included students in junior high and high school. On the other hand, with the exception of Amaral et al. (2002) and Lynch et al. (2005), studies with ELL samples included only students in the upper elementary grades (i.e., 3rd and 4th grade).

Designs. Studies with non-ELL samples tended to use short, multi-week science units in which specific writing strategies were used to enhance instruction (e.g., Akkus et al., 2007; Keys et al., 1999; Patterson, 2001). Studies with ELL samples, on the other hand, used longer, often multi-year designs with multiple literacy and professional development interventions (e.g., Amaral et al., 2002; Lee et al., 2005; Lee et al., 2008). At the same time, studies with non-ELL students included quantitative, qualitative, and mixed-method designs while studies with ELL samples were all quantitative.

Discussion and suggestions. The contrasts in research context and design between the groups of studies elicit an obvious question: Why the marked difference between studies with non-ELL and studies with ELL samples? The answer likely centers on research needs and purpose. In the United States, the urgency to raise student science scores – especially minority student scores (National Center for Education Statistics, 2011) – is conducive to large-scale and multifaceted interventions. On the other hand, studies in international contexts repeatedly center their purpose on testing writing to learn theories: does writing in the science classroom fact improve learning? (See for example: Hand et al., 2006) or how does writing enhance science inquiry instructions? (See for example: Akkus et al., 2007). The purpose of studies outside of ELL contexts is not on an urgency to raise student scores, but to enhance learning and explore the role of writing in that process.

Not surprisingly, studies with non-ELL samples include qualitative and quantitative studies, while studies with ELL samples are only quantitative studies. Funding involving critical at-risk, diverse populations favors large-scale quantitative studies with good reason: high-stakes assessments are the current measuring stick for student achievement on a national level and quantitative studies are most conducive to numerical results which can be more easily compared to standardized assessments (Kieffer et al., 2009).

Research related to science and writing, however, could intersect design types and grade levels. Larger scale intervention studies could be conducted with non-ELL populations and focused writing interventions and/or analysis in science with ELL students could be carried out. At the same time, research in writing in science with younger students at the mainstream/ international level, for example, and with ELL students at the junior high and high school level in the United States should also be considered.

Non-ELL and ELL study interventions and measurements compared. As has been noted, studies with non-ELL samples focus on specific writing strategies and studies with ELL samples focus on multifaceted interventions that included writing. What kinds of writing interventions, however, did researchers specifically use in these two categories and how was student conceptual understanding measured?

Writing interventions. Perhaps the clearest way to break up the kind of writing interventions used in studies with non-ELL samples is to think of them as *structured* and *unstructured* writing interventions regardless of whether the intervention used conventional writing strategies, creative/personal writing-to-learn strategies, or a mix of both. Structured writing interventions are interventions in which students are explicitly taught a writing strategy. From the literature review, it is clear that structured interventions yield positive results for students' conceptual understanding while unstructured do not. Examples of structured writing interventions are the Science Writing Heuristic (SWH) model used in Akkus et al.'s (2007) study and the specific writing tasks students were instructed in (i.e., were given models and practice) in Hand et al.'s (2009) study. Unstructured writing interventions are interventions in which the teacher did not provide students with an explicit writing strategy. An example of an unstructured writing intervention is the one used in Mason's (2001) study in which students were asked to write before, during, and after a science activity but were not explicitly told to follow a certain writing structure or strategy.

Studies with ELL samples in large research contexts were more difficult to classify in terms of writing intervention types. The researchers provided brief descriptions of how writing was used in the classroom, likely because writing was not the sole focus of their large-scale intervention studies. Amaral et al. (2002), for example, mentioned how students were expected to “collect, record, analyze, and report data,” (p. 224) in their notebooks – a process that may have been structured (but not necessarily so). Fradd et al. (2001) and Lee et al. (2005) noted how students wrote science expository and narrative paragraphs – a process that may have been less structured (but not necessarily so). Lee et al. (2008) mentioned using writing in the science classroom intervention and Lynch et al. (2005) mention using science notebooks. Because the descriptions of writing use in the science classroom within the larger studies conducted with ELL students are limited, classifying and/or drawing conclusions about which kinds of writing interventions are useful for promoting the conceptual understanding of students or how writing specifically impacted conceptual understanding of students is not possible. It is possible, however, to infer that writing played a role – along with other variables – in the increase of ELL conceptual understanding.

Measurements. Studies with non-ELL samples varied in their measurements of student conceptual understanding in accordance with their overall research design. Qualitative studies in this category, by nature, included analysis of the writing process students engaged in during the short interventions and looking for patterns and evidence of student understanding in their writing (Keys et al., 1999; Patterson, 2001; Mason, 2000). The quantitative studies (Akkus et al., 2007; Gunel et al., 2009; Ritchie et al., 2011; Hand et al., 2009) and the mixed-method studies (Mason & Boscoallo, 2000; Rivard & Straw, 2000) in this category, all used analysis of covariance (ANOVA) or some variation of the statistical method (i.e., ANCOVA or MANCOVA) to compare writing mean scores from pre to post-test as a measure of student conceptual development progress.

Researchers working with ELL samples used either national and state standardized assessments or district/researcher-developed assessments to measure students' conceptual understanding over time (i.e., conceptual development). Amaral et al. (2001) used the Stanford Achievement Tests and state standardized tests for reading, math, and writing, for example; while Fradd et al. (2001), Lee et al. (2005) used district developed science tests. Lee et al. (2008) and Lynch et al. (2005) used researcher-developed pre and posttests.

Discussion and suggestions. It is noteworthy that many of the studies in both categories came from the same group of researchers. In the non-ELL category, researchers invested in the Science Writing Heuristic (i.e., Gunel, Hand, and Keys) were part of four out of the nine studies in the area of science and writing (44%). In the ELL category, Lee and colleagues constituted three out of the five studies (60%). While the fact that certain groups of researchers focus on a research area is not a negative thing, there is certainly room for other research groups to step in and perhaps even transcend the ELL and non-ELL boundaries. What would happen, for example, if structured writing intervention such as the SWH were used with ELLs in the context of science inquiry? What would happen if science inquiry with writing integration studies were conducted with mainstream populations, not just in the United States? What would the outcomes of the SWH be on younger students? Does student conceptual understanding of science benefit from writing regardless of context, design, intervention, and method?

A glimpse into the answer to the latter questions lies in Lynch et al.'s (2005) quasi-experimental study where ELLs did not show growth in conceptual development after using a science and literacy intervention while non-ELL students did improve. Clearly, science inquiry interventions can work for non-ELL students, but this sole study calls into question the appropriateness of the measurement – a point that is noteworthy given that measurement of conceptual understanding is inherently tied to language (Kieffer et al., 2009). The kinds of instruments used to measure conceptual understanding should consistently and thoroughly be described and/or developed according to the population and their linguistic backgrounds. Mixed-method studies designs where researchers approach writing in science from both a qualitative and quantitative point of view, as Mason and Boscollo (2000) did, could also be a beneficial approach to understanding the role of writing in students' conceptual understanding of science.

Last but not least, it is clear that students benefit conceptually by having strong instructional scaffold in place, regardless of the type of writing interventions (conventional, creative, mixed) in science. Educators should continue to provide strong scaffolds in terms of lesson design and instructional deliver/modeling regardless of the strategy use. Studies that compare outcomes between conventional, creative/personal, and/or a mix of writing-to-learn strategies on students conceptual understanding are also worth conducting and exploring.

Writing and Academic Language

Studies with non-ELL (N = 2) and ELL (N = 3) samples in this category were both few in number (Total: N = 5). Overall, researchers reported positive findings on students' language development due to writing in science, though not all results were positive (Ruiz-Primo et al., 2004) and the ELL studies needed time to show improvement (Amaral et al., 2002; Lee et al, 2005; Lee et al., 2009). The following discusses and compares the studies' contexts, designs, interventions, and measurements and provides suggestions for future study.

Contexts and designs. Only two studies with non-ELL samples looked specifically at language development in the science classroom and one was conducted in Southern Wales (Patterson, 2001) while the other was conducted in the United States (Ruiz-Primo et al., 2004). As has been noted, Patterson (2001) noted positive language growth while Ruiz-Primo et al. (2004) did not note positive language growth. The two studies differed greatly in design. Patterson's (2001) study was qualitative and included a structured writing intervention while Ruiz-Primo et al.'s (2004) study was quantitative and did not include structured writing instruction as part of the intervention. At the same time, in Patterson's (2001) study, the researcher was involved in the instructional intervention whereas in Ruiz-Primo et al.'s (2004) study, the researchers approached data analysis in an archival manner, looking at writing samples after the fact, and creating a rubric to determine the writing quality as well as to determine the kinds of instructional practices that took part in the classroom (which, they found deficient due to lack of teacher feedback in writing). It is thus difficult to parallel the two studies or draw any conclusions from the miniscule sample. It is interesting, however, to consider how different research approaches and contexts can lead to such disparate results in terms of language development for students.

Studies with ELL samples that considered language development were also few (N = 3); so again, conclusions cannot be definitely drawn. Observations of the studies, however, can inform future research. All of the studies were conducted in the United States; all of the studies were quantitative – one was post-hoc (Amaral et al., 2002) and two were pre-posttest (Lee et al., 2005; Lee et al., 2009); and all of the studies found gains in language development for ELLs over time. The latter is an important observation in terms of second language development theory amongst ELL populations given that L2 theory notes that second language acquisition needs time to develop (Cummins, 1981).

Interventions and measurements. All studies except Patterson's (2001), which was qualitative and focused on a specific writing intervention, had writing interventions embedded within larger science-inquiry instruction with samples of ELL students. The one exception was Ruiz-Primo et al.'s (2004) study, which did not specify a sample including ELLs but was drawn from students who participated in a large-scale science intervention study in the United States. The fact that studies like Patterson's (2001) focused on writing interventions did not look at language development is, at first, startling. As has been previously discussed, however, the purpose of the studies in international contexts with specific writing interventions was, overall, to note how writing as a tool helped to enhance science concept learning (not language learning). From this angle, it makes sense that the majority of the studies interested in language development would fall into the category of ELL students in the United States. This is not to say, however, that studies with non-ELL samples should consider language development alongside science development, especially if science language and concept are believed to be related.

Measurements for language development varied according to the overall research design (i.e., qualitative and quantitative). Patterson's (2001) qualitative study, for example, counted the number of words students used to measure fluency (i.e., more words = more fluency) and the number of connective words to measure language complexity (i.e., more connective words = more effective explanations). Quantitative studies used different researcher-created rubrics to quantify language development based on analysis of student writing samples and then perform statistical analysis (Lee et al., 2005; Lee et al., 2009; Ruiz-Primo et al., 2004). Of these studies, Ruiz-Primo et al.'s (2004) stands out from the other studies in that one of the study's main goals was to further validate their science notebook writing rubric.

Suggestions. Studies in which literacy activities are integrated into the curriculum *and* track student's language development are clearly needed with populations of non-ELL and ELL students. As researchers and theorists in the science field have noted, science constitutes its own language (Gee, 2005; Lemke, 1990) and can be especially challenging for ELL (Ryoo, 2010) and, arguably, low-SES students. For this reason, researchers measuring literacy integration in the science classroom should make an effort to consider measures of language development alongside measures of science conceptual understanding. At the same time, researchers should continue to strive for quality measurements of language development and conceptual understanding in the context of science writing – not an easy feat but one that can build upon studies such as Ruiz-Primo's et al. (2004).

Relationship between Academic Language and Conceptual Understanding

Studies in this category were few in number (N = 5). The researchers in these studies all agreed that some relationship between language and concept exists even though their studies varied in context, design, interventions, and measurement, as will be discussed below.

Contexts & designs. All of the studies included samples of either middle school (Keys et al., 1999; Keys, 2000) or high school (Gunel et al., 2009; Ruiz-Primo et al., 2010) students, with the exception of one study that had a sample of fifth grade students (Ruiz-Primo, 2004). Notably, all of the studies were conducted in the United States; yet, none of the researchers used samples from populations with ELL students.

The qualitative studies came from the same researcher (Keys) and two out of the three quantitative studies came from Ruiz-Primo's work. Again, this fact is not negative in itself, but it does point to the need for more researchers to analyze the relationship between language and concept reflected in science writing. Interestingly, all of the studies included relatively small samples; even the quantitative studies selected random writing samples from the larger sample, likely to keep the analysis manageable.

Interventions and measurements. Studies in this category resembled those in the first category (i.e., *Writing and Conceptual Understanding*) in that two described structured writing interventions (Keys, 2000; Gunel et al. 2009) while three described unstructured writing interventions (Keys et al., 1999; Ruiz-Primo et al., 2010; Ruiz-Primo et al., 2004). Furthermore, studies with structured writing interventions reported more promising findings than studies with unstructured writing interventions. An important commonality within this group of studies is that each of them mentioned the importance of the quality of writing instruction on the part of the teacher. As the researchers implied, in order for language to develop and reflect science conceptual understanding and, conversely, for language to aid in the development of science understanding, writing in science must be taught well.

Measurements between the studies varied. Key's qualitative studies used functional grammar analysis (1999) and linguistic and thematic classification of writing samples (2000) in order to establish links between language and conceptual understanding. Quantitative studies used researcher-created rubrics to quantify student writing and then used the writing scores for regression analysis (Gunel et al., 2009) to predict science scores on posttest multiple-choice measures and correlational analysis within the writing (i.e., understanding and language scores) (Ruiz-Primo et al., 2004) and outside the writing (i.e., with other forms of science assessment) (Ruiz-Primo, 2004; Ruiz-Primo et al., 2010).

Suggestions. Studies that explore the relationship between language and concept should include samples from the lower grades (i.e., K – 5th) as well as samples with ELL learners. The latter is a major research gap. Studies in the context of science and literacy integration with ELL students claim to be fundamentally interested in the conceptual and language development of ELLs and even low-SES students (i.e., Amaral et al.; Lee et al., 2005; Lee et al., 2009). Therefore studies with ELL and low-SES samples should consider the linguistic and conceptual development of students due to science and writing integration interventions *and* the relationship between language and concept. The application of correlational analysis such as the kind that Ruiz-Primo (2004; 2010) conducted can and should be conducted with samples of writing from ELL and low-SES students. In this way research can continue to confirm and unravel theories regarding the role of language in science learning.

Student Language Status Groups

Researchers have compared statistical gains in science concept scores (Amaral et al., 2002; Lee et al., 2005; Lee et al., 2008; Lee et al., 2009; Lynch et al., 2005) and academic language scores (Amaral et al., Lee et al., 2005; Lee et al., 2009) between ELL and non-ELL students in the context of science inquiry interventions. Two studies (Lee et al. 2005; Lynch et al., 2005) have compared science concept score differences among ELL, former ELL, and non-ELL students. No studies have compared language scores among ELL, former ELL, and non-ELL students. The following discusses the researchers' findings and their implications.

Comparing language and concept scores in science education between ELL and non-ELL students is logical given the achievement gap that exists between the two groups (NCES, 2011) and the overall goal for research to find interventions to close the achievement gap (Lee, 2005). Studies comparing these two language status groups consistently found that ELL students, though making comparable gains in science and academic language achievement scores over time, scored significantly lower than non-ELL students in language (e.g., Lee et al., 2005) and concept scores (e.g., Lee et al., 2008; Lee et al. 2005, Lynch et al., 2005).

Researchers that compared ELL, former ELL, and non-ELL students (Lee et al., 2005, Lynch et al., 2005) within the context of science inquiry interventions looked at *concept* scores (not language) based on pre-posttest standardized assessments. Lee et al. (2005) reported that ELL and former-ELL students scored significantly lower than non-ELL students at posttest. Lynch et al. (2005), on the other hand, reported that ELL students scored significantly lower than former ELL and non-ELL students. It is not possible, therefore, to draw definite conclusions from previous work with respect to former ELL students given that only two studies do this and given that their findings differ. Comparing language and concept scores between ELL, former ELL, and non-ELL, therefore, is noteworthy because so little research exists that tracks former ELL students, in the context of science intervention studies.

What does exist regarding ELL, former ELL, and non-ELL comparisons outside of science intervention studies tends to be situated in the context of assessment studies (e.g., Abella, 2005) and attitude studies (e.g., Lindholm-Leary, 2001). Yet, former ELLs should be a point of interest because students who have been exited from bilingual and ELL programs should theoretically possess enough academic language to achieve as well and non-ELL students. For example, Lindholm-Leary (2001) found that former

ELLs were more likely to highly rate quality bilingual programs over non-ELL and ELL students. Former ELLs likely felt more success as a result of the program. At the same time, former ELL students may or may not possess CALP if exited too early (Cummins, 1981); therefore, research working with samples including former ELL students can measure their achievement in their analysis. For example, Abella (2005) found former ELLs to have difficulty exhibiting their content-area knowledge on math achievement tests, possibly because of “language and cultural barriers” (p. 127).

Unfortunately, the one study (Lee et al., 2009) considering science writing concept and language scores with a sample of ELL, former ELL, and non-ELL students research in the context of science-literacy integration – in order to perform HLM analysis – lumped non-ELL and former ELL students together in their analysis. Lee et al. (2009) did, however, state how “Further research could…test whether the relationship between writing form [i.e., language] and content [i.e., conceptual understanding] differs by English proficiency” (p. 166). Studies including ELL students in their sample population should therefore continue to look at differences in science and language and concept achievement between student language classifications – including former ELLs – as well as explore relationships between language and concept across students groups.

Gender and Science Writing

Researchers in five out of the nineteen studies from the literature synthesis on science and writing examined gender as a variable. The following breaks down research findings based on whether the studies included ELLs in their sample or not.

Studies with non-ELL samples. Researchers (Patterson, 2001, Ritchie et al., 2011, Rivard & Straw, 2000) in three out of the thirteen studies with non-ELL samples considered gender as a variable. Boys had higher scores in measures of science recall (Rivard & Straw, 2000) and science interest (Ritchie et al., 2011) than girls. One researcher (Patterson, 2001) found girls preferred rigid planning structures for science writing while boys preferred the less prescriptive concept mapping structure of the writing intervention presented in the study. Notably, Ruiz-Primo et al. (2004) - the one study considering science notebook scores - did not consider gender as a variable.

Studies with ELL samples. Researchers in two (Lee et al., 2005; Lynch et al., 2005) out of six studies with ELL samples considered gender as a variable. In their analysis of pre-post test scores, researchers in both studies found no significant difference between science scores for girls and boys on language and concept scores at the end of one academic school year.

Gender and Science Achievement

Given the above findings, the following will briefly synthesize past research on science achievement that considers gender, race, and SES status as a variable. Doing so provides a context for the research on writing in science and informs the current study.

Gender. Researchers (Bacharach, Baumester, & Furr, 2003; Jones, Mullis, Raizen, Weiss, & Weston, 1992; Mullis, Dossey, Owen, & Phillips, 1993) note girls perform lower in science achievement than boys, with difference more noticeable in high school than in middle school. In their quantitative study using data from the National Educational Longitudinal study with a sample of 8th – 12th grade students, for example, Bacharach et al. (2003) found the average yearly increase in science achievement for boys was larger than average yearly increase for girls.

Gender and race. Race, however, seems to play a more marked role in science achievement differences between girls and boys (Jones, Mullis, Raizen, Weiss, & Weston, 1992) with racial differences in science achievement measures appearing much earlier than gender differences in science achievement measures (Mullis, Dossey, Owen, & Phillips, 1993). Furthermore, race differences account for more variance in science achievement throughout all grades than gender differences do (Hanson, 1996). For example, in Bacharach et al.'s (2003) more recent study discussed above, the researchers classified their sample of 8th – 12th grade students as White and Black and found racial differences to account for a greater disparity between achievement over time than gender.

Gender, race, and SES status. Socio-economic status (SES status), however, affects students' science achievement regardless of race or gender. For example, using longitudinal data from the National Education Longitudinal Study (NELS: 88), Muller, Stage, and Kinzie (2001) found SES status among the variables that strongly and positively correlated to students' eighth-grade achievement across all races and gender subgroups. Kohlhaas, Lin, and Chu (2010) examined the relationships among gender, ethnicity (i.e., race), and poverty (i.e., SES status) with fifth graders' (*n* = 8,741) science performance. The researchers examined fifth grade data files (2003–2004), from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), a nationwide study. Again, while all three of their variables (i.e., gender, race, and SES status), contributed significantly to fifth graders' science performance, differences existed on each main effect. As the researchers put it so succinctly, "The smallest to the largest mean differences between subgroups are gender (3 points), poverty [SES status] (14 points), and ethnicity [race] (16 points)" (p. 8). Thus, race and SES status account for more mean difference than gender.

Importance of gender as a variable. The fact that race and SES status play a larger role than gender in science achievement differences between boys and girls, however, does not discount the fact that gender does play some role in discrepancies between male and female students. Researchers intersecting fields of education and sociology (Chen, 2009; Correll, 2001) continue to ask why a majority of college ready males choose science, math, and technology careers over females. Some researchers (Riegle-Crumb, Moore, & Ramos-Wada, 2011) argue that the cumulative effect of gender disparities over time, beginning as early as junior high, has an effect in career choice for students by the time they are ready to go to college. Moreover, as has been noted, few science writing intervention studies (N = 2) with ELL students in their sample consider gender as a variable. Including the examination of gender as a variable in studies considering science achievement is clearly needed, especially in studies including samples of linguistically diverse and low-SES students. This study will therefore consider gender as a variable of comparison in the students' science notebook scores.

Science Notebook Writing Rubrics

From the literature review, few researchers invested in rigorously considering rating instruments for writing. Seven researchers mention using writing rubrics to quantify science writing (Amaral et al., 2002; Gunel et al., 2009; Lee et al., 2005; Lee et al., 2009; Rivard & Straw, 2000; Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010). However, only three researchers provide details on rubric development, reliability, *and* validation (Rivard & Straw, 2000; Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010), likely because the studies are mostly embedded within larger interventions (e.g., Lee et al., 2005). To clarify, “details on rubric development” means more than just reporting “inter-rater reliability” or that a “rubric was developed”. Rigorous rubric development and reporting entails details about expert reviewers, reliability estimates, and perhaps rigorous content and/or construct validity. For example, in the studies providing less detail, science writing rubrics reliability estimates are reported, but the rubric is simply reported to either be researcher-created based on state standard rubrics (e.g., Amaral et al., 2002; Lee et al., 2005; Lee et al., 2009) or teacher-created (Gunel et al., 2009). In the more detailed studies (Rivard & Straw, 2000; Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010), the researchers cite theories to back up their constructs, and expert reviewers to help with validation, in addition to reliability estimates.

If research instruments are to account for accurate measurement of the construct at hand, they must be critically evaluated as being rigorous. For example, Liu, Lee, and Linn (2011) discussed the critical need to develop scoring rubrics for constructed-response items on science exams (i.e., response items that require students to write) in order to accurately and fairly capture students' responses. The researchers stressed the importance of "valid, authentic, and efficient assessments" in science education (p. 1079), the need to focus on valid constructed-response rubrics for ELL students due to the "language demands [constructed-response items] place upon the test taker" (p. 1084), and the need to establish strong inter-rater reliability when dealing with constructed-response items. The same concerns and principles apply to science notebook entries, which are, in essence constructed-response items in the context of real classroom use.

Quantifying science notebook writing entries with a writing rubric is challenging due to the varying form of the entries. Science notebook entries can include illustrations, lists, graphs, tables, figures, and embedded diagrams with text labels. As has been noted, from the literature review only one group of researchers reported details on how they developed and calibrated the reliability of a rubric specifically for measuring science notebook entries. Ruiz-Primo et al. (2004) created a multifaceted rubric in which raters first identified the kind of writing task – or "genre" – the students' were writing and then rating the students "quality of writing" based upon how well their writing aligned to the task, or genre. The researchers based their logic on the theory that language is meant to accomplish a communicative purpose, and genres reflect the purpose of writing (Lemke, 1990). There is certainly to develop, build upon, elaborate, and strengthen science writing rubrics in order to ensure a more accurate and precise measurement of science writing.

Conclusion

If the achievement gap in science achievement is to be closed and overall avoided amongst students, especially amongst students who are ELL and low-SES (NCES, 2011), research must consider instructional interventions that foster students' scientific literacy. As has been discussed, one way to do this is to focus on two of the major components that encompass scientific literacy: science conceptual understanding and academic language (Kieffer et al., 2009; Lee, 2005). A powerful tool for fostering student conceptual understanding and academic language in science classrooms is writing (Rivard & Straw, 2004; Yore, 2003).

The findings of this literature synthesis lead to several conclusions. The first is that writing can impact science conceptual understanding for all students, whether the student is ELL, former ELL, or non-ELL. However, studies that include ELL students are embedded in large-scale intervention studies with multiple variables (e.g., Amaral et al., 2002; Lee et al. 2005), making it difficult to decipher whether or how much of student conceptual understanding and/or growth was specifically due to the writing intervention *unless* the study specifically analyzed student writing as Lee et al. did in their 2009 study. Studies with non-ELL students, on the other hand, focus specifically on writing interventions in the science classroom (e.g., Akkus et al., 2007; Mason & Boscollo, 2000), making it more plausible that the writing intervention impacted students' conceptual understanding and/or growth. Successful writing-to-learn interventions, however, were contingent upon whether or not scaffolds were in place to aid students' understanding as they engaged through the writing process, whether the writing was conventional, creative, or a mix of both.

The second conclusion is that writing in science has the potential to improve students' academic language. Researchers working with ELL populations reported student academic language improvement over time (Amaral et al., 2002; Lee et al., 2005; Lee et al., 2009). The two studies with non-ELL students (Patterson, 2001; Ruiz-Primo et al., 2004) had mixed findings. Again, the findings are not certain given the small literature sample size. The third conclusion is that a relationship between language and concept may exist, given that teachers provide proper science writing instruction (Gunel et al., 2009; Keys et al., 1999; Keys, 2000; Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010).

Last, researchers who attempt to analyze and/or quantify science notebook writing and/or their effect on student outcomes are rare (Amaral et al., 2002; Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010). The rarity of these studies is likely due to the difficulty of defining exactly what kind of writing genres are being produced in the notebooks (See Ruiz-Primo et al., 2004) as well as the fact that science notebooks seem to take a secondary role in students' learning of science if simply integrated into the curriculum instead of seen as a primary and important tool that teachers should scaffold (Amaral et al., 2002). In addition, only Amaral et al. (2002) look at science notebook writing in the context of a sample of ELL students.

Study possibilities in the realm of science and writing are many. As has been touched upon, researchers in international contexts could apply larger-scale science inquiry and literacy integrated interventions. Researchers in the United States with ELL students could conduct studies with more focused and structured writing interventions. Researchers considering writing in science could consistently include analysis of science understanding and language development *and* explore correlations between the two constructs. In addition, there is room for the development and/or use of robust writing measurement instruments like the rubrics used in Ruiz-Primo et al. (2004)'s study, especially for use with ELL and low-SES students who struggle with the acquisition of science literacy. Studies that observe and quantify the quality of science and literacy instruction are also needed in order to draw definite conclusions regarding the effectiveness of student development due to specific interventions in the science classroom. A gap exists in research that explores both the academic language and conceptual understanding of students reflected in science writing and the intersection of both constructs with populations of ELL and low-SES students. Studies considering the effect of science notebooks, especially with populations of ELL students, need to be conducted and explored. Finally, variable such as language classification groups, which include former ELLs and gender as a variable can and should be explored in order to provide a more comprehensive picture of individual variables that could affect students' performance as well as to provide much needed insight to the existing, but small body of literature in this area.

In conclusion, academic language and conceptual understanding, which are part of what make up scientific literacy, are critical for students' success in science (Lee, 2005), are the key elements needed for this group of students to succeed in science achievement tests (Kieffer et al., 2009), and can be fostered in the context of science inquiry with writing integration (Amaral et al., 2002). Researchers, therefore, need to measure language development and conceptual understanding as reflected in students' science writing as well as explore how language and concept are related for ELL and low-SES students. Finally, researchers including ELL students in their studies should continue to look at differences between student language classifications and gender, in order to verify differences in achievement.

CHAPTER III

METHODOLOGY

The purpose of the present study was to (a) investigate the impact of science notebook writing on students' academic language development over time, across student language status and gender groups (b) compare students' conceptual understanding across student language status and gender groups, and (c) explore the extent of relationship between students' academic language and conceptual understanding scores.

This chapter lays out the methodological design of the study. The chapter includes sampling, research design, context of the study, program intervention, instrumentation, data collection, data analysis, and a summary.

Sampling

The present study derived from Project Middle School Science for English Language Learners (MSSELL) (NSF Award No. DRL - 0822343), a two-year (2009 - 2011) federally funded project that targeted approximately 270 ELL and non-ELL non-ELL diverse students from low-SES backgrounds in a large urban school district in Southeast Texas. The objective of the study was to implement a rigorous, two-year randomized trial longitudinal evaluation of enhanced science instruction for middle school students (grades 5 and 6) whose first language was Spanish; however, the study also included minority students of low-SES backgrounds who were integrated in the classrooms and whose first language was English. The hypothesis of the larger study was that a research-based model of science instruction would improve science achievement and academic English proficiency.

The overall study had an experimental design at the school level and a quasi-experimental design at the student level. Four out of the ten intermediate schools (grades 5-8) in the selected school district were randomly assigned to either a treatment or control conditions with the school administrator's permission. As a result, two schools were assigned the treatment condition and received enhanced science practice and two schools were assigned the control and received typical science practice.

For the present study, only treatment students who participated in the first year of the intervention, receiving enhanced science practice, were considered for sampling purposes for the writing analysis (n = 210; average age of 12.40 years, *SD* = .66). This is because science notebooks were required and collected only from the treatment condition.

An *a priori* test was conducted in order to determine the appropriate sample size for statistical significance at p = .05 using the G*Power analysis online software (Faul, Erdfelder, Buchner, & Lange, 2009). For a Pearson r correlation analysis, the minimal sample size was 19 (effect size of .70); for a one-way, 3 level ANOVA, the minimal sample size was 24 (effect size of .80); and for a mixed within-between ANOVA, the minimum sample size was 30 (effect size of .80). The final sample size of two sets of 90 mean scores (one set for language and one set for concept) discussed below was therefore sufficient for the present study.

Research Design

A stratified, random sample of 30 students was drawn from a pool of 210 students who participated in the first year of the treatment condition of the overall study. The sample was stratified at the student classification level (i.e., ELL, former ELL, and non-ELL) to ensure an equal number of students fit each group category for ANOVA analysis purposes. In addition, the sample was stratified at the gender level to ensure equal representation in each group.

Within each student's science notebook the following three units, each representing 1-2 weeks of instruction, were rated: (1) Physical Science Unit, (2) Earth/Space Science Unit, and (3) Life Science Unit. The units were chosen because they represented student work from the beginning, middle, and end of the academic year, with approximately 10 weeks separating the unit's implementation. Each unit included an average of 6 entries, and each student's notebook had the possibility of a maximum of 18 individual writing entries to be scored (this is due to the fact that students may have been absent and missed and entry). The maximum number of pages to be rated was therefore 540 (18 x 30).

Each notebook included two sets of three "mean scores" (i.e., the sum scores divided by the number of entries identified in each students' notebook) and two sets of "grand mean scores" (i.e., the sum of the mean scores divided by 3; 3 representing each unit). The first set of three "mean scores" represented the student's language score for each unit. The second set of three "mean scores" represented the student's concept score for each unit. The first set of "grand mean scores" represented the student's overall language score for the notebook, and the second set of “grand mean scores” represented the student’s overall concept score for the notebook. As Ruiz-Primo et al. (2004) noted, using the “mean scores” with the same scale (from 1-4), allows for comparison of student performance on different aspects (in the case of the present study, language and concept).

The number of notebooks was kept small in order to keep the analysis manageable and to allow for quality of analysis, while still accounting for the power needed to attain statistical significance for a balanced ANOVA design (*n* needed = 24), a mixed between-within ANOVA (*n* needed = 30), and a correlational analysis (*n* needed = 19). The number of language mean scores was 90 (30 notebooks x 3) and the number of concept mean scores was 90 (30 notebooks x 3). Table 1 illustrates the break down of the sample students whose writing samples were selected from the intervention schools.

Table 1

Sample Break Down for One Set of Mean Scores

	ELL	Former ELL	Non-ELL
Male	5 x 3 = 15	5 x 3 = 15	5 x 3 = 15
Female	5 x 3 = 15	5 x 3 = 15	5 x 3 = 15
Total	10 x 3 = 30	10 x 3 = 30	10 x 3 = 30

Note. Each student accounted for a total of two sets of 3 mean scores (one for language and one for concept), each representing approximately 6 pages of a science unit, totaling approximately 540 unique writing entries.

As Table 1 indicates, an equal number of ELL, former ELL and non-ELL students were selected from the intervention classrooms. In addition, an equal number of males and females represent each group. It should also be noted that all of the students were classified as low-SES and that each student accounted for a total of two sets of three mean scores (one for language and one for concept), making the final writing sample size 90 for each set of scores (i.e., language scores and concept scores), with the scores comprising of a total of approximately 540 individually rated writing samples.

Context of the Study

The present study took place in a large urban school district in Southeast Texas in which 66.9% of the students were classified as Hispanic and 28.3% were classified as African American (TEA, 2010). Furthermore, 85% of the students qualified for free or reduced lunch, which is an indicator of low-SES status (TEA, 2010). The district was chosen because it had many years of experience working with ELLs and low-SES students, a consistent philosophy in terms of instructional implementation, ease of access to regular and ELL programs within the district, and a reputation for academic excellence.

Program Intervention

The overall program intervention consisted of two components:
(a) professional development, and (b) enhanced science instruction. The components are discussed as follows.

Professional Development

Teachers attended approximately 18 bi-weekly meetings totaling approximately 42 hours of training (initial training and then bi-weekly training). During the training sessions, research coordinators provided three-hour training sessions on topics including English as a second language (ESL) strategies; assessment of teaching practice; teacher reflection; and inquiry lesson practice. The components of the training are discussed below.

ESL strategies were based on the work of Herrell and Jordan (2008). The ESL strategies included the following: (a) using realia (i.e., authentic materials such as real newspaper articles) and manipulatives (i.e., concrete objects that allow students to explore using hands-on approaches to learning) to help make concepts concrete for students during lessons; (b) integrating other content areas into science instruction to promote cross-content understanding and skill re-enforcement; (c) integrating technology such as software and smart boards into lessons to increase student engagement; (d) cooperative learning in which students worked together to complete tasks and/or discuss questions posed to promote oral language use and problem-solving skills; (e) advanced organizers (i.e., graphics by which words and/or objects can be arranged to promote conceptual understanding of the relationship between ideas); (f) visual scaffolding (i.e., using images and words that can be seen and heard to promote comprehensible input of information); and (f) questioning (i.e., asking questions that promote higher-level thinking – *how* and *why* vs. *what* questions - for students).

The assessment of teaching practice training sessions allowed teachers to provide feedback on how classroom lessons went in the classroom. As part of enhanced science instruction, teachers received structured lesson plans aligned to national, state, and English language proficiency standards. Lesson plans were written in English, followed the 5-E model structure for inquiry lessons (Bybee, et al., 2006), were scripted, and included sections for first language (i.e., Spanish) clarification of concepts for ELL students (A sample lesson plan is provided in Table 3). Teachers shared ideas on what worked for them in the lessons, and the trainers provided suggestions. Teachers also received feedback on classroom observations conducted as part of the fidelity of implementation.

Teacher reflection sessions followed the Reflection Cycle (Brown & Irby, 2000), which utilized artifacts from the teachers' science classroom and asked them to reflect, describe, and appraise the events associated with the artifact. Teachers were then asked to transform their behaviors based on the reflection and were asked to complete surveys and reflective entries in teacher portfolios.

During inquiry lesson practice sessions, teachers were trained to follow the projects' scripted lesson plans. If inquiry activities were included, then the activity was outlined step by step in the lesson plan. During the training session, teachers had a chance to explore the materials and do the activities themselves before presenting the inquiry lessons to the students.

Enhanced Science Instruction

The following explains components of the enhanced science instruction the intervention students received. Each component is described in detail, with research citations to justify inclusion of the component.

Science inquiry. Teachers delivered daily 85-minute science lessons that followed the 5-E model of instruction. The 5-E model of instruction is an inquiry-based format for creating science lesson plans. The 5-E model aims to encourage discovery and higher-level thinking for students (Bybee et al., 2006) as well as to scaffold teachers in thoughtfully structuring and delivering science lessons to maximize student learning within the inquiry model. Each lesson in the intervention was designed to cover all of the 5-E components in one class period to the extent possible, given the time. Table 2 breaks down the components of the 5-E model. Each stage is named and described through an illustration of what the teacher and students are expected to do at each stage of the inquiry sequence.

Table 2

The 5-E Model

Stage	What the Teacher Does	What the Student Does
Engage	Creates interest; focuses student thinking; raises questions; allows students to make connections between past and present learning.	Ask questions and show interest in the topic.
Explore	Provides an environment in which students work together to manipulate materials, explore, and problem solve; observes students; asks probing questions; acts as a consultant.	Tests and forms predictions and hypothesis; discusses with others; asks questions; records observations and thoughts.
Explain	Encourages students to explain concepts on their own; formally provides definitions and explanations; uses students previous experience for explaining concepts.	Explains possible answers; listens to others' explanations; uses recorded observations when explaining.

Table 2 Continued

Stage	What the Teacher Does	What the Student Does
Elaborate	Encourages students to apply or extend concepts they learned in new situations; expects students to use formal definitions and explanations previously learned.	Practice skills and/or learning more information in order to develop deeper and broader understandings; records observations and explanations.
Evaluate	Observes students as they apply new concepts and skills; assesses students' knowledge, skills, and understanding of concepts.	Answers questions using previously learned knowledge; asks questions to prompt future investigations; evaluates own progress.

Note. Adapted from Bybee et al. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications executive summary.* Colorado Springs, CO: Biological Sciences Curriculum Study.

Table 3 illustrates a scripted lesson plan from Project MSSELL following the 5-E model.

Table 3

Sample MSSELL Scripted Lesson Plan Following 5-E Model

Week 5, Day 4, 85 minutes

DOWLS: Science Probe p. 25 #2 [Pocket Folders] **(7-10 minutes)**
Every student receives a half sheet for warm-up. Pass out individual copies to students.
Display warm-up on ELMO.
Read prompt with the students. Students identify the type of energy produced by each object and explain whether matter is a solid.
Let students discuss their responses with their partners.
Call on students randomly to share their responses.

Engage: Mystery Matter **(5 minutes)**
Place metal sphere inside the box the night before. Show students the box and explain that you have placed matter in the box and you want to know what it is.
Questions: ***Explain how we can identify it as a solid, liquid, or gas.***
Identify another physical property we can test.
Refer to the objective for the day:
Identify the type of force a magnet demonstrates.
Predict how we can test matter to see if it is attracted to a magnet

Table 3 Continued

Week 5, Day 4, 85 minutes
Explore: Magnetic Test [Journal] **(20 minutes)** 1. Students copy chart into their journal. 2. Discuss materials (one baggie/group of steel paper clip, plastic spoon, aluminum foil, plastic toy, wax candle, glass marble, penny) with students and have them copy them into the materials column. 3. Students record their predictions. 4. Students then test all materials and record results. Students write responses to the following in their journal with the chart: ***Identify the objects that were attracted to the magnet. Explain how you would classify these objects according to your data from your chart.*** ***Draw a conclusion about what type of matter is attracted to a magnet.***
Evaluate/ Product: Critical Thinking Question 5.7A #1 [Journal] **(7-10 minutes)** Display question on the ELMO. Read the passage with the students. Give students time to think about their responses. Students record their responses in their journal. Students write their explanation to justify their responses. Let students discuss their responses with their partners. Call on students randomly to share their responses. Display and discuss student work sample on the ELMO.
Closure: (3 minutes) ***Identify the physical property that we used to classify matter.***
Review Homework (15 minutes) Display homework on the ELMO. Walk student through the process of how to work through the problems using strategies. Relate the problems back to the activities and investigations from the week.

Note. Words in italics are scripts the teachers were to follow. In the original lesson plans, the words were printed in blue and bolded; the "Engage" portion of the 5-E lesson was not used daily, but was used when appropriate to the lesson and time available.

The science lessons were embedded within science units aligned to the Texas Essential Knowledge and Skills (TEA, 2011c). The TEKS are state standards aligned with national standards for each content area, delineating what should be taught at each grade level, including critical vocabulary words. In science, the TEKS include topics in four major science categories and content standards that must be taught: physical science, earth/space science, and life science. Within each category, teachers design lesson sequences as long as they cover the content objectives and critical vocabulary. An excerpt of the scope and sequence for the first year (Grade 5) of the project MSSELL lesson plans, including the break down of the science units, can be found in Appendix C.

Questioning. Science inquiry instruction stresses the role of questioning on the part of students and the teacher to promote higher-level thinking. Questioning strategies, in fact, were found to have the highest effect size on student achievement based on a meta-analysis of U.S. research published from 1980 to 2004 (Schroeder, Scott, Tolson, Huang, & Lee, 2007). For this reason, teachers in the intervention group modeled specific strategies and implemented answering techniques meant to promote student questioning (i.e., timed thinking, visual cues, choral response, pair-share, quick write).

Direct instruction. As a way to support science inquiry discussion and understanding, teachers provided direct vocabulary instruction. Direct vocabulary instruction has been found to be effective in building students' vocabulary (e.g., August, Carlo, Dressler, & Snow, 2005). Structures were set up to ensure teachers modeled academic science language and students were able to use it included presenting student-friendly definitions, using visual scaffolding, and having students create science journal glossaries.

Literacy integration. As has been noted, integrating literacy into the content areas is an effective and critical component for scaffolding students towards scientific literacy (e.g., Fang, 2004, 2006). Increasing scientific literacy leads to science understanding and achievement for students (e.g., Fang, 2006; Lee, 2005; Yore, 2003).

Reading. In order to re-enforce science concepts (Fang, 2006), students engaged in a structured reading practice of expository science texts. Before reading, the teacher introduced critical vocabulary words students would then partner-read (for fluency and comprehension), and then pairs would ask each other scripted questions (for comprehension). Critical vocabulary words, based on the TEKS, were defined in the unit plans (See Appendix C for an example of critical vocabulary words embedded in the units). The class would then come back together to review and clarify possible misconceptions.

Writing. Students kept individual science notebooks. Science notebook entries aided students in processing and solidifying science concepts as part of the 5-E inquiry lesson model (Butler & Nesbit, 2008). Writing tasks included (a) recording vocabulary words and definitions in a glossary; (b) illustrating and labeling diagrams; (c) organizing information using two-dimensional figures; (d) recording observations and predictions; and (e) reflecting on field trips and/or writing perspective-based entries (i.e., newspaper article formats). Teachers also received training on providing feedback (grammar and content-related) on the students' content and language.

Other Components

The following discusses other components of the intervention designed to enhance instruction and learning in the classroom.

Technology integration. Effective technology integration has been noted to be a positive teaching practice for promoting ELLs' learning (Waxman, 2002) and science inquiry learning (Kim & Hannafin, 2011). Treatment classrooms, therefore, used technology hardware such as projectors, interactive whiteboards, document cameras, digital cameras. Science-based educational software, and internet resources were also used in the classroom.

Family involvement. Studies have advocated the importance of family involvement for the academic success of students, especially students of diverse (Panferov, 2010) and low-SES backgrounds (Manz, Fantuzzo, & Power, 2004). Students in the treatment classroom received take-home booklets related to the science topic units being studied in the classrooms. The booklets were written in English and Spanish and included elements such as fun facts, science activities that could be done in the home, extra readings, brief assessments, and crossword puzzles. A short letter to the family introducing the topic and a parent signature page were also included for accountability. Teachers held one 45 minute meeting in the fall of the academic year to explain to parents and guardian how to use the take-home booklet.

Instrumentation

Two rating rubrics were used to rate each student's (1) science language quality (or, for brevity, *language*) and (2) science conceptual understanding (or, for brevity, *concept*) as reflected in the science notebook entries for three different time points (beginning, middle, and end of the year). The rubrics were adapted from Ruiz-Primo et al.'s (2004) study in which the researchers developed and tested rating scales that could reliably produce scores for what the researchers called "quality of communication" and "understanding" (p. 1483) given any type of science notebook entry. [1]

---

[1] In addition, the Ruiz-Primo et al. (2004) study sought to quantify the student's "opportunity to learn" (p. 1478) by evaluating the amount of teacher feedback found in the science notebooks. The present study deals only with the constructs of language and conceptual understanding.

Rubric Rational: Literature

Ruiz-Primo and colleagues have worked on formative assessment research, including science notebooks, with publications dating back to the 1990's. The researchers have published in reputable journal articles such as the *Journal of Educational Measurement*, the *Journal of Research in Science Teaching*; for nationally funded center such as the *National Center for Research on Evaluation, Standard, and Student Testing* (CRESST) at Stanford University; and major educational research conferences such as the American Educational Research Association (AERA) (e.g., Li, Ruiz-Primo, Ayala, & Shavelson, 2000; Ruiz-Primo, Baxter, & Shavelson, 1993; Ruiz-Primo, Li, Ayala, & Shavelson, 1999; Ruiz-Primo, Li, Ayala, & Shavelson, 2004; Ruiz-Primo, Li, & Shavelson, 2002; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Ruiz-Primo, Li, Tsai, & Scheneider, 2010, Ruiz-Primo, Li, Tsai, & Schneider, 2010). From the literature review, it is clear that no other group of researchers created rigorous rubrics specifically for rating *science notebook entries*. Furthermore, the rubric used in Ruiz-Primo et al.'s 2004 study was developed specifically to rate 5th grade science notebooks used within the context of science inquiry units – the same purpose of the present study.

Rubric Rational: Psychometrics

The researchers cite two main studies in which they piloted and then used the science notebook rubrics, which the present study adapts (Ruiz-Primo et al., 1999; Ruiz-Primo et al., 2004). Both studies report high inter-rater reliability and decent validity measures.

Ruiz-Primo, Li, Ayala, and Shavelson's (1999) pilot study of the science notebook rubric also used in their 2004 study, reported an average inter-rater reliability score of 0.86 for the language component of the rubric, and 0.88 for the conceptual component of the rubric. The rubric's validity was established by correlating science notebook scores on two units to student unit performance scores for each unit (0.65 effect size for the "Variable Unit" and 1.30 effect size for the "Mixtures Unit").

The researcher's 2004 study, on which the rubric for the present study is based upon, reported an average inter-rater reliability score of 0.82 for the language component of their rubric and 0.86 for the conceptual component of their rubric. The rubric's validity was established by correlating students' language and conceptual scores. The researchers noted positive correlations between the language and conceptual notebook scores with magnitudes of 0.53 in the "Variables Unit" and 0.52 in the "Mixtures Unit".

Unfortunately, the Ruiz-Primo et al. studies (1999; 2004) do not specify the type of inter-rater reliability analysis used, and it assumed that the researchers used percent-agreement since this method is mentioned in Table 5 of their 2004 study. The critical detail of the statistical method used to determine inter-rater reliability is addressed in the present study.[2]

Rubric Rational Summary

Given the above reasons (i.e., no other group of researchers have created rubrics to quantify science notebooks, that the 2004 study of Ruiz-Primo et al. was created for a

---

[2] Note: Ruiz-Primo, Li, Tsai, and Schnider's (2010) study reports using generalizability theory to report reliability coefficients for a rubric specific to rating students' scientific explanations in lab reports as does the Ruiz-Primo, Baxter, & Shavelson (1993) study.

very similar purpose and for a similar population as the purpose and population of the present study), I used the rubrics in the Ruiz-Primo et al. (2004) study as a basis from which to create rubrics that would quantify the notebook entries for the present study. The present study's instrumentation specifics, adaptation and reliability measures are discussed as follow.

Language Rubric

The purpose of the language rubric was to rate how the students groups (ELL, formal ELL, and non-ELL) developed their academic language over time *and* compared in their academic language scores. All of the writing samples for the present study were written in English. The following explains the constructs that comprise the language rubric and the adaptations from Ruiz-Primo et al.'s (2004) science notebook rubric.

Construct 1A: Quality of communication. Ruiz-Primo et al.'s (2004) rubric viewed "quality of communication" as a reflection of the students' ability to use academic language appropriately in the context of specific science writing tasks.[3] In their rubric, separate scoring criteria were defined by the general characteristics of each task. For example, the task of "defining" focused on using academic vocabulary and using verbs in the present tense to explain the meaning of a word. The task of "illustrating and labeling diagrams", on the other hand, focused presenting identifiable information with the appropriate labels and technical (i.e., academic) language. The

---

[3] Note: the researchers use the term "genre" which appears synonymous to the science notebook entry *tasks* – that is, whether the students were writing definitions, reflections, summarizing findings, predicting, hypothesizing, or simply describing an observation or procedure or reporting findings. To avoid confusion and for consistency, the term "task" will be used rather than "genre" in the present study.

present study adopts this definition of "quality of communication" as the basis for quantifying student science notebook entries.

In Ruiz-Primo et al.'s study (2004) the researchers identified 14 tasks reflected in science notebook entries (See: Ruiz-Primo et al., 2004 for a published list of the tasks). In the present study, the larger study noted the following five tasks were used within the science notebook writing intervention: (a) defining within the notebook (b) illustrating and labeling diagrams; (c) organizing information using two-dimensional figures (i.e., charts); (d) recording observations and predictions; and (e) reflecting. Ruiz-Primo et al.'s (2004) categories, which matched the tasks, were therefore used to create the rubric for the present study.

Construct 1B: Conventions of communication. Inevitably, the quality of communication is linked to the students' use of academic language *as well as* the students' ability to express ideas clearly (i.e., grammar, spelling, and syntax) within the given task. While Ruiz-Primo et al.'s (2004) rubric accounted for the academic language of science as used appropriately within a given science notebook entry task, it did not account for the mechanics (i.e., punctuation, spelling) and grammar (i.e., word order, syntax) of the English language that are critical for clear communication. As Lee et al. (2009) noted, "English proficiency involves knowledge and effective use of linguistic skills, including phonemes, syllables, morphemes, vocabulary, grammar (syntax), and written conventions (e.g., punctuation, capitalization, spelling)" (p. 154). Given that the present study includes ELL, former ELL and non-ELL students from low-socioeconomic backgrounds (who may have English registers different from the standard or academic registers of English), the rubric for the present study includes criteria specifically for measuring conventions of communication (i.e., grammar and mechanics in English).

The *conventions of communication* portion of the rubric was adapted from portions of the Texas English Language Proficiency Assessment System (TELPAS) writing rubric for ELL's in 2nd – 12th grade (TELPAS, 2011). The TELPAS was created by the Texas Education Agency in order to fulfill federal requirements for assessing the English language proficiency of ELLs in K-12 grade in the four language domains: listening, speaking, reading, and writing. The present study adapted a portion of the TELPAS writing rubric, which is also comprised of a four-point scale, to include criteria for rating students' grammar and mechanics in English.

Summary of construct 1. In summary, the language quality portion of the rubric integrated two aspects of language: quality of communication and conventions of communication. The aspects were combined within one rubric on the theoretical premise that language – especially academic language in science – encompasses technical vocabulary, discourse patterns specific to science tasks (Gee, 2005; Lemke, 1990) and grammatical features inherent to communication (Lee et al., 2009; Scarcella, 2003). In practical application, a student's science notebook entry may contain much writing but be so unclear that it is difficult to decipher whether the scientific task is addressed at all or may include little writing that is nonetheless precise and clearly aligned to the task at hand. The complete, adapted language rubric can be found in Appendix D.

Rating process. For the present study, raters scored the quality of communication for each task on a four-point scale. The four-point scale was meant to go beyond general categories such as “completeness, clarity, and organization” in addition to distinctly focusing on task characteristics vs. functional analysis such as lexical density and clause characteristics (Ruiz-Primo et al., 2004, p. 1484). Rating according to task on a four-point scale for Ruiz-Primo et al.’s study (2004) and the present study allowed for a more sensitive rating scale.

Language rubric use. As Ruiz-Primo et al. (2004) noted, the ability to use academic language (i.e., quality of writing) should theoretically improve over time, so the present study used mean scores as a measure of academic language growth over time in addition to comparing scores across student groups. Given that all students in the present study, regardless of English language level, were expected to engage in the academic language of science in English, the samples were rated with the same rubric in order to measure the academic language of the student in the context of English science instruction.

Concept Rubric

The purpose of the concept rubric was to rate how the students groups (ELL, formal-ELL, and non-ELL) compared in their science conceptual understanding on three science units. The following explains the break down of the conceptual rubric, adapted from Ruiz-Primo et al. (2004).

Construct 2: Writing as a reflection of understanding. Writing in science can theoretically serve as a process *leading to* conceptual understanding (Yore, 2003) as well as a *demonstration of* conceptual understanding (Halliday & Martin, 1993). The science writing entries in this study were approached as *demonstrations of* science conceptual understanding, as Ruiz-Primo et al. (2004) did. This definition is in line with that National Science Education Standards (National Research Council, 1996) definition of conceptual understanding as being able to use science knowledge. This approach does not to discount that writing is, itself procedural; but for the present study, the writing entries are taken as a demonstration of student understanding at a given time point.

In the present study, the entries from three science units were used in the rating process: physical science, earth/space science, and life science. State standards (TEA, 2011), based on the National Science Standards (National Research Council, 1996), define the concepts students should understand within each science topic. In turn, schools use the standards to create specific lesson objectives, or defined concept objectives. For a list of the state standards in place at the time of the intervention and the defined concept objectives for each unit rated in the science journals, see Appendix C which contains the unit science unit lesson plans which align to notebook entries rated. These state standard and defined concept objectives were used to train raters to use the conceptual rubric and to rate the science notebook entries.

Rating process. Following Ruiz-Primo et al.'s (2004) rubric structure, the concept rubric was used to rate conceptual understanding on a four-point scale. The rating was, like in the language rubric, based on the entry type (i.e., defining within the notebook, illustrating and labeling diagrams, organizing information using two-dimensional figures (i.e., charts), recording observations and predictions, and reflecting).

Concept rating rubric adaptation and use. Studies measuring conceptual *development* in science as reflected in writing have used standardized assessments as a comparison point (e.g., Amaral et al., 2002) or pre-posttest assessments measuring the same, specific concept (e.g., Lee et al., 2008; Ritchie et al., 2011). However, studies like Ruiz-Primo et al.'s (2004) and the present study dealt with archived documents representing a variety of units and concepts. It is therefore not logical not measure conceptual *development* within science notebooks because each science unit is measuring a different set of concepts.

In addition, not all concepts are equally easy for students to grasp. In science in particular, state data shows that 5th grade students have the most difficulty with earth science concepts, then physical science concepts, and finally life science concepts with some variation (J. Jackson, Personal Communication, August 30, 2012).

Last, a rubric for conceptual understanding must fit the content of the unit being analyzed. Therefore, the present study used the sample rubric provided in Ruiz-Primo's et al.'s (2004) study as a model for rating the three units from the sample. Mean concept for each unit score, under the assumption that each unit was testing the same overall concept, were used to compare scores across student categories (ELL, former ELL, non-ELL) and gender groups (male, female) but *not* across time (i.e., development). The complete concept rubric can be found in Appendix D.

Rubric Advantage

As has been noted, the instrument for the present study was adapted from the Ruiz-Primo et al. (2004) study. The advantages of rating rubrics for quantifying writing and for statistical analysis are two-fold: (a) the rating scales compared language for any entry, so the entries could be compared equally regardless of content and (b) the rating scales for the language and content rubric included word anchor antonyms, so the scores produced interval data, necessary for ANOVA and correlational statistical analysis (Thompson, 2008). However, because the instrument was *adapted*, the training, reliability, and validity of the instrument must be addressed.

Validity

Analysis for the present study's rubric focused on both content and construct validity. The following will describe each validation analysis.

Content validity. Content validity addresses the following: If an instrument is appropriate for its intended purpose, experts should agree that the instrument will measure what it claims to measure (Huck, 2008; Reynolds, Livingston, & Willson, 2006). In addressing content validity, the present study adds a dimension missing from published studies that either do not address content validity (i.e., Lee et al., 2009; Ruiz-Primo et al., 2004). A rubric, however, should be a potentially useful instrument at the research and classroom level in which the science notebooks are used as assessments of students' learning, not just as a measure for a single study.

To address the content validity of the rubrics, I sent the language and content rubrics to four expert reviewers. One reviewer was a nationally recognized professor of science education specializing in science literacy with ELL students and amply published in the field of science education with ELL students; one reviewer was a professor of science education actively involved in teacher training (in science and ELL instruction) and science curriculum development within the state in which the study took place and familiar with the state science standards; one participant was a science curriculum specialist in the school district in which the intervention took place with thirteen years of ELL teaching experience and seven years experience in curriculum development; and one participant was a former science teacher of ELL middle school students in the state in which the intervention took place.

The reviewers provided feedback and suggestions on the rubric to ensure that it was assessing the academic language and conceptual understanding of the students. In specific, the participants addressed (a) whether the rubrics were appropriate to measuring the construct (i.e., academic language of science; conceptual understanding of science) (b) concerns about the rubrics (c) specific suggestions on how to improve the rubrics, if any. Comments, concerns, and suggestions regarding the rubrics are summarized in Table 4.

Table 4

Rubric Feedback

	Comments	**Concerns**	**Suggestions**
Professor 1	Structural components of 1, 2, 3, etc. seem reasonable	Difficult to understand overall	Communicate theory or construct
Professor 2	Concept rubric is good	Difficult to understand language rubric without examples; Also difficult to separate "language from concept"	Define things such as "technical terms" with specific TEKS
Curriculum Specialist	With some revision, rubrics will be useful tools for authentic assessment of ELLs	Some descriptors are vague. (i.e., in language rubric what is "frequent, occasional, and minimal"; in concept rubric, what is "logical justification"?)	Consider defining/listing examples

Table 4 Continued

	Comments	Concerns	Suggestions
Educator	Rubrics are both appropriate for the construct	Scoring for punctuation is unclear	Define what punctuation will be counted
Summary	Rubric structure and potential is positive; concept rubric is easier to understand than language rubric	Rubrics are unclear in some sections (especially language rubric), making understanding difficult	Rubric needs definitions and examples

As can be noted from Table 4, Professor 1's concerns were the most general (i.e., "difficult to understand"), Professor 2's concerns and the curriculum specialists' concerns were more specific (i.e., "what is logical justification?"), and the Educator's concerns were narrowly focused on the issue of how to grade punctuation. Common themes emerged in terms of a general concern of lack of clarity and a need for the rubric to define and provide examples for the ratings which are included in the summary column – namely that the language rubric is overall more difficult to follow and the rubrics in general need more specific definitions, examples, and explicit alignment to the standards.

Rubric refinement. Overall, the feedback pointed to the fact that the rubric was standing alone and out of context when sent to the reviewers. In other words, trained raters would have access to specific examples and to state, district, and classroom lesson standards. However, it was clear that the rubric as a stand-alone instrument could be significantly improved. As a result, the language and concept rubric were refined by doing the following: (a) Providing specific examples within the rubrics where appropriate (i.e., defining), (b) Adding bullet points to make language rubric more user friendly/easier to read, and (c) Creating a “Rubric Manual” for raters to refer to with specific technical terms and concepts (See Appendix E).

The issue of specifying frequency of errors did not seem logical as ELL mistakes are difficult to quantify in terms of number. The wording for this section of the rubric was taken from a state developed and validated rubric specifically made for ELL students (TELPAS, 2011). Therefore, the wording was not changed; however the latter part of the criteria, which describes how much the errors hinder or do not hinder understanding, was emphasized by underlining the text. As has been noted, the science notebook writing rubric can be found in Appendix D.

Construct validity. Construct validity integrates a collection of evidence to determine the degree of validity of the interpretation of scores on some measure (Huck, 2008; Reynolds, Livingston, & Willson, 2006). To address the construct validity of the rubrics, I adopted an internal approach to construct validity which assumed the following: If scores on language and concept reflect students’ achievement in scientific literacy, then the language and concept scores should correlated with each other.

Therefore, if language and concept score are related, scores should have positive correlations. If, however, language and concept are not measuring aspects of scientific literacy, then the correlation should not be positive and its magnitude should not be high. Table 5 illustrates the results of a Pearson *r* correlation between overall language and concept scores broken down by content domains. (Note: Correlations broken down by language status groups and gender are presented in the Results section of this study).

Table 5

Correlations between Language and Concept Mean Scores

Domain	Pearson *r*
1. Physical Science	.843** (n = 26)
2. Earth/Space Science	.878** (n = 26)
3. Life Science	.810** (n = 26)

** Correlation is significant at the .01 level (2-tailed).

As Table 5 indicates, the correlations between language and concept scores were all positive and large. Overall, the results indicate that the language and concept scores were tapping into similar aspects of scientific literacy and thus internally validating the rubric. Correlations were largest for Domain 2 (Earth/Space Science), which had the least amount of writing/language, and smallest for Domain 3 (Life Science), which had the most amount of writing/language. Perhaps as language demands increase, the measures begin to tap more into slightly different aspects of scientific literacy.

Reliability

For the present study, I randomly selected 20% (*n* = 6) of the total notebook sample (*n* = 30) to train raters and to calculate reliability, given previous research calibrates with 20% of the total sample (e.g., Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010). I then used one notebook (approximately 18 pages) to train and calibrate raters across the three science domains (i.e., physical science, space science, earth science, and life science). The raters consisted of two other doctoral students and myself, who were all working toward doctoral degrees in bilingual education. After training, the three raters scored the remaining five notebooks (approximately 72 pages) to calculate reliability estimates.

G theory. Inter-rater reliability was calculated using Generalizability Theory (G theory), a method that uses analysis of variance to estimate variation due to sources of error beyond just measurements over time (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Advantages of using G theory over other estimates of reliability such as percent-agreement include G theory's ability to account for more sources of error and provide estimates of the magnitude of variance so that the greatest measurement error can be pointed out (Raykov & Marcoulides, 2011). Within G theory, a G-study provides a generalizability coefficient (alpha) as well as the degree of variance of interactions between variables. If need be, a D-study can be conducted to allow the researcher to decide which facets can be removed for efficiency while still maintaining a high generalizability coefficient or which facets need to be added to increase the generalizability coefficient. I conducted both a G-study and a D-study for each of the rubrics (language and concept) described as follows.

G-studies. For the G-studies, I used a full factorial design for both the language and concept rubric. A full factorial design is the ideal G-study design because it allows partitioning the variance components (i.e., person effect, rater x person interaction effect, etc.) in order to estimate (alpha) *all* interactions. For analysis of both the language and concept rubric, Facet 1 was defined as the three raters (i.e., doctoral student 1, 2, and 3) and Facet 2 was defined as the three science domains (i.e., physical science, earth/space science, and life science) based off of the national and state standards (National Research Council, 1996; TEA, 2011c). Both facets were defined as random facets, rather than fixed, meaning that the raters and science domains were pulled from a pool of all possible raters and all possible science domain concepts. Table 6 is a visual of the full factorial design of the G-study for both the language and concept rubrics.

Table 6

G-Study Design for Language and Concept Rubrics

	Domain 1			Domain 2			Domain 3		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
Person 1									
Person 2									
Person 3									
Person 4									
Person 5									

Note. “Person” indicates the notebook entries of each of the 5 students.

The G study reliability estimates, percent agreement (as a point of comparison),[4] and Ruiz-Primo et al.’s (2004) reliability estimates (as a point of reference for our adaptation) are compared visually in Table 7.

---

[4] Since we had 3 raters, we calculated percent agreement by taking the mean level of agreement across all pairs of reviewers.

Table 7

Comparison of Reliability Estimates

	Ruiz-Primo et al. (2004): type of reliability estimate not clear	Present Study: Generalizability Coefficient	Present Study: Percent-Agreement
Language Rubric	0.82	0.893 (0.017)	0.936
Concept Rubric	0.88	0.858 (0.022)	0.920

Note. Values inside the parenthesis are the relative error of the generalizability coefficient.

In the present study, I calculated inter-rater reliability using G theory across 3 raters, using 5 notebooks. Ruiz-Primo et al.'s (2004) study calculated inter-rater reliability across 3 raters using 24 notebooks. Unfortunately, the Ruiz-Primo et al. (2004) study does not clearly indicate what reliability estimate was used. Still, as Table 7 indicates, the generalizability coefficient for the language rubric was 0.893, with a relative error of 0.017. The coefficient is higher than the reliability estimate of the language rubric (0.82) reported in the Ruiz-Primo, et al. (2004) study. The generalizability coefficient for the concept rubric was 0.858, with a relative error of 0.022. The coefficient is slightly lower than the reliability estimate of the concept rubric (0.88) reported in the Ruiz-Primo, et al. (2004) study. Not surprisingly, percent-agreement was higher for the language (0.936) and concept (0.920) rubric. Given G-theory accounts for more variance than percent agreement, and in comparison to the Ruiz-Primo et al. (2004) study, the results of the present study indicate high reliability for the present study's rubric using G theory.

D-studies. D-studies were not necessarily needed but were conducted in order to analyze where the highest variance fell for the rubrics and to hypothetically see which facets could be reduced for future studies using the rubrics. For the D-studies for the language and concept rubric, I first determined which of the variables accounted for most of the error for each of the rubrics. In this case, because the generalizability coefficient was so high, the D-study focused on which variable component it could reduce while still maintain a high generalizability coefficient (vs. which variable component to add in order to increase the generalizability coefficient). The variance estimates for the language and concept rubric are listed in Table 8.

Table 8

Variance Estimates for the Language and Concept Rubric

Component	Estimate for language	Estimate for concept
Var(Person)	.142	.133
Var(Domain)	.071	.036
Var(Rater)	.005	.002
Var(Person*Domain)	.017	.026
Var(Person*Rater)	.007	.013
Var(Domain*Rater)	-.002	.012
Var(Person*Domain*Rater)	.082	.078
Var(Error)	.000	.000

As can be noted in Table 8, *domain* accounted for most of the error for the language and concept rubric (apart from person, which is common and a variable that cannot be controlled post-hoc). The number of domains, however, could not be reduced because all three time points for the physical, earth/space, and life science were needed to measure language growth over time and were desirable for comparing content domain scores. The number of raters, though, could be reduced. Calculations showed the generalizability coefficient based on using two raters instead of three raters was 0.860 for the language rubric and .826 for the concept rubric, still high coefficients. Future studies could consider using two instead of three raters.

For the present study, all three raters were utilized for the sake of efficiency (i.e., it was faster to use three raters and all of them were available to rate). The remaining 24 notebooks were randomly distributed among the 3 raters. Each rater rated 8 notebooks, equaling approximately 432 pages to obtain the language and a concept scores for each notebook.

Data Collection

Researchers on site collected student science journals from the intervention schools at the end of the first year of the science intervention in the Spring of 2010. From this pool, the main research coordinator on site collected a stratified, random sample. I then chose three entries – one from the beginning, one from the middle, and one from the end – of the science journal at approximately equal intervals according to the academic school year (early Fall 2009, early Spring 2010, late Spring 2010). Trained raters coded and rated the samples and recorded the data on recording sheets (See Appendix F), which I then imputed into a database. I took part in the data entry and analysis of the science writing entries. On the one hand, this fact poses as a bias threat to the internal validity of the study. On the other hand, the possible bias threat is mitigated by the fact that I was not involved in intervention process – rather, the analysis was meant to be objectively analyzed "after the fact." Keeping the possible bias in mind, however, I ensured that my ratings were reliable and in line with my colleagues' ratings by, as has been discussed, implementing a rigorous reliability study.

Data Analysis

Researchers who quantify science writing favor ANOVA for statistical analysis in order to compare the mean results between and within groups of variables (e.g., Lee et al., 2005; Lee et al., 2009; Ruiz-Primo et al., 2004). Mixed between-within subjects ANOVA, an extension of repeated measures ANOVA, is a robust and efficient statistical analysis researchers can use when planning to measure both between-subjects variables and within-subject variables over time (Pallant, 2010). In addition, Pearson *r* correlation analysis, used to describe the direction and strength of the relationship between two interval variables (Pallant, 2010; Thompson, 2008) can be used to analyze the relationship between the two dependent variables (e.g., language and concept scores). For the present study, three main analyses were used to answer the research questions of which one used mixed between-within subjects ANOVA, one used ANOVA, and one used a Pearson *r* correlation. To test hypothesis, therefore, the present study used statistical software SPSS. To meet the assumption requirements for applying ANOVA and mixed between-within subjects ANOVA, data were analyzed. In specific, descriptive statistics, results of tests of homogeneity of variance and sphericity, interaction effects, effect sizes, and visual representations of the data are reported.

Research Question 1: Academic Language Development

A mixed between-within subjects ANOVA was run to answer research question 1: Over the course of 1 year, did ELLs, former ELLs, and non-ELL/low-SES 5th grade students make significant gains in academic language, and to what extent does the level of academic language across student language status (ELL, former ELL, non-ELL) and gender groups differ? Using a mixed between-within subjects ANOVA analysis, I first looked at the within-subjects variable of time for each groups' language score (i.e., the between-subject variable). I then compared the between-subject variable, language development, across student groups (i.e., ELL, former ELL, non-ELL and gender).

Research Question 2: Conceptual Understanding

Three separate 3 (i.e., language status) x 2 (i.e., gender) ANOVA analyses were run (one for each conceptual domain) to answer research question 2: To what extent does the level of conceptual understanding across student language status (ELL, former ELL, non-ELL) and gender groups for science notebook entries differ? In each 3 x 2 ANOVA analysis the between-subject variable, conceptual understanding, was compared across the same student groups (i.e., ELL, former ELL, non-ELL and gender).

Research Question 3: Relationship

Three Pearson *r* correlational analysis were run in order to answer research question 3: To what extent are academic language and conceptual understanding related as reflected in science notebook entry scores, and how do the relationships compare between student language status (ELL, former ELL, non-ELL) and gender groups? Correlation coefficients were used to interpret strength of correlations between language and concept scores for each group as well as to descriptively compare correlations between groups.

Summary

Chapter III of my study includes a detailed description of the research design, data collection, and analysis methods. Researchers analyzed and recorded writing scores from three time points from the student science notebooks. The results of the data analysis are presented in the following chapter.

CHAPTER IV

RESULTS

This chapter presents the results of the data analysis to answer each research question. Descriptive statistics, results of tests of homogeneity of variance and sphericity, interaction effects, effect sizes, and visual representations of the data are reported accordingly.

Language Development

Descriptive statistics of language science notebook scores from the beginning (Domain 1), middle (Domain 2), and end (Domain 3) of the year (2009-2010) for the language status groups are listed in Table 9 and for the gender groups in Table 10.

Table 9

Descriptive Statistics for Three Time Points (Language Status Groups)

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ELL							
D1	7	2.1243	.37206	-1.018	.794	-.375	1.587
D2	7	2.2600	.55311	1.501	.794	1.704	1.587
D3	7	2.8586	.63302	1.096	.794	.571	1.587
Former ELL							
D1	9	2.4578	.42005	-.376	.717	-1.402	1.400
D2	9	2.6078	.35209	-.287	.717	-1.813	1.400
D3	9	3.0356	.44722	1.341	.717	1.890	1.400
Non-ELL							
D1	10	2.6310	.37817	.687	.687	-.323	1.334
D2	10	2.9120	.31808	.687	.687	-.980	1.334
D3	10	2.8730	.43987	.687	.687	-1.153	1.334

Note: D1 = Domain 1 (beginning of year), D2 = Domain 2 (middle of year), D3 = Domain 3 (end of year)

Table 10

Descriptive Statistics for Three Time Points (Gender Groups)

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Female							
D1	14	2.6207	.24408	.203	.597	-1.164	1.154
D2	14	2.7379	.37326	-.623	.597	-.078	1.154
D3	14	2.9407	.46892	.622	.597	.613	1.154
Male							
D1	12	2.2175	.50171	.482	.637	-1.086	1.232
D2	12	2.5142	.55400	.307	.637	-1.303	1.232
D3	12	2.9075	.52640	.425	.637	.303	1.232

Note: D1 = Domain 1 (beginning of year), D2 = Domain 2 (middle of year), D3 = Domain 3 (end of year)

According to Table 9, in total, 7 ELL students, 8 former ELL students, and 10 non-ELL students had scores for the three time points; therefore the ANOVA was very nearly balanced, especially between the former ELL and non-ELL group. According to Table 10, 14 female and 12 male scores had scores for each of the three time points; once again, the ANOVA was very nearly balanced for the gender groups. Table 9 and 10 indicate the absolute values of skewness and kurtosis statistics are between +/- 1 and +/- 2 for both the language status groups and the gender groups, acceptable values for psychometric purposes (Cutting, 2012), which underlie ANOVA assumptions (Thompson, 2008). This means the data are normally distributed.

The statistics also indicate that as time progressed, the means for the ELL, former ELL, and both gender groups increased progressively across the three time points. Means for the non-ELL group increase from Domain 1 to Domain 2 but dipped in Domain 3 (though remained higher than in Domain 1). Means for the female group were higher overall than for the male group, but means for the female and male groups got closer together as time progressed.

Having established the data are normally distributed, a mixed between-within subjects ANOVA was conducted to assess the impact of a science notebook writing intervention on students language status (ELL, former ELL, and non-ELL) and gender groups' academic language across three time periods (Domain 1 [beginning of year], Domain 2 [middle of year], and Domain 3 [end of year]).

Homogeneity of variance assumptions for the data, though not perfectly met at the univariate level, were met at the multivariate level. Leven's Test of Equality of Error Variance, a univariate measure, yielded the value for Domain 1 as .032, Domain 2 as .046, and Domain 3 as .317; thus, two of the values were statistically significant beyond .05. Non-statistical significance is desired to meet the assumption, however researchers note the robustness of the mixed between-within ANOVA compensates when not every value is non-statistically significant at the univariate level (Dickinson, 2011). Box's Text of Equality of Covariances Matrices, a multivariate measure, was non-statistically significant at the .001 level with a value of .080, indicating the homogeneity of variance assumption was not violated at the multivariate level. Therefore, the analysis could proceed.

Data met the sphericity assumption. Mauchly's Test of Sphericity yielded a value of .224, indicating non-statistical significance, meaning the sphericity assumption was not violated. This assumption is important in repeated measure designs because it means that all pairs of levels of the within-subjects variable have equivalent correlations

(Tabachnick & Fidell, 2007). Had the assumption of sphericity been violated, Type 1 would have been increased and the *p*-values would not accurately reflect the observed statistics, giving a biased picture of the results (Keselman, Algina, & Kowalchuck, 2001).

Interactions between time and language status group [Wilks' Lambda = .756, *F* (4, 38) = 1.426, *p* = .244, partial eta squared = .130], time and gender [Wilks' Lambda = .804, *F* (2, 19) = 2.323, *p* = .125, partial eta squared = .196], and time, language status group and gender, [Wilks' Lambda = .849, *F* (4, 38) = .811, *p* = .526, partial eta squared = .079], were not statistically significant. A substantial[5] main effect for time did exist [Wilks' Lambda = .396, *F* (2, 19) = 14.514, *p* = .000, partial eta squared = .604], with all groups increasing in language scores overall across the three time periods. The main effects comparing the language status groups [*F* (2, 20) = 4.194, *p* = .030, partial eta squared = .295] and gender groups [*F* (1, 20) = 4.773, *p* = .041, partial eta squared = .193], were statistically significant and fairly substantial.

A Tukey post-hoc test, appropriate for comparisons interested in simple contrasts and with smaller sample sizes in order to maintain power (Thompson, 2008), was run to determine where the differences existed between the language status groups. Mean differences between the ELL and former ELL group (*p* = .155, std. error = .14553) and between the former ELL and non-ELL group (*p* = .713; std. error = .13268) were not statistically significant. Mean differences between the ELL and non-ELL group were

---

[5] Partial eta squared, or the effect size magnitude is being gaged according to the commonly used guidelines proposed by Cohen (1988, pp. 284-287): .01=small effect, .06=moderate effect, .14=large effect.

statistically significant (*p* = .034; std. error = .14321). Figure 1 illustrates differences between the groups at each of the three time points.

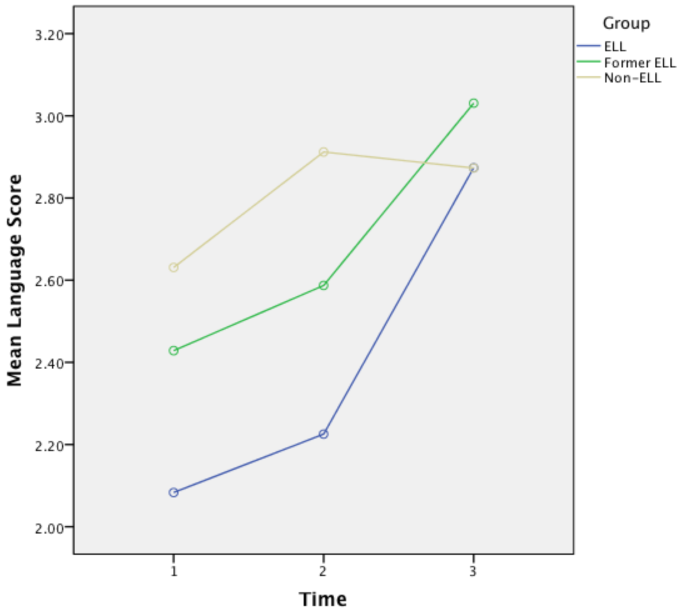

Figure 1. Differences between language status group language scores over time.

Figure 1 illustrates the differences in means noted in the descriptive statistics. As can be noted in Figure 1, ELL and former ELL groups' language scores increased over the three time points and the non ELL group increased from the first to second time point, dipping at the third time point, but still remained higher than the first time point. A marginally significant interaction effect (*p* = .090) was noted for time and language status group when considering tests of within subject contrasts as repeated measures. Tests of within-subject contrasts, specifically noted a marginally statistically significant value (*p* = .080) from time point 2 to time point 3. This means the amount of growth in students' language scores from time point 2 to time point 3 was affected by their language status. From Figure 1, in appears ELLs made the most growth, followed by former ELLs and, last, by non-ELLs.

Mean score differences between genders were statistically significant and substantial (*p* = .041, partial eta squared = .193), as previously noted. Figure 2 illustrates the differences.

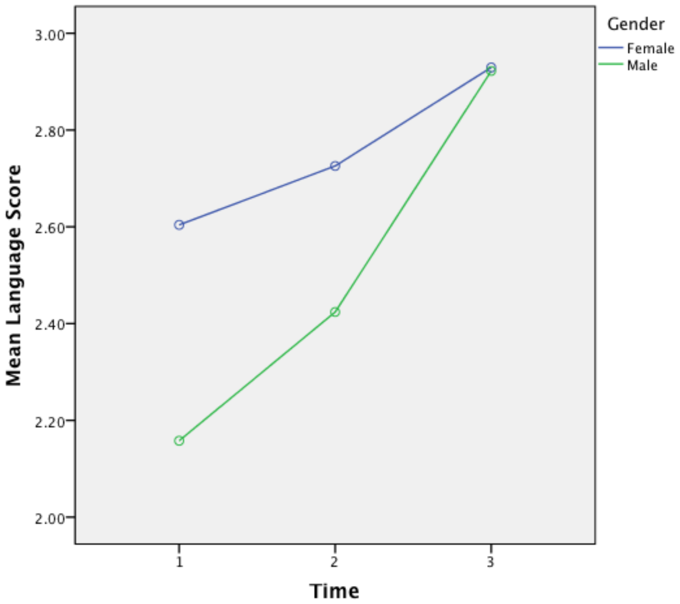

Figure 2. Differences between gender group language scores over time.

Figure 2 illustrates the differences in means noted in the descriptive statistics, with both females and males increasing in language scores over the three time points, females having a higher mean for the first two time points, and males catching up to the females by the third time point. No significant interaction effects were noted between time and gender groups when considering tests of within subject contrasts as repeated measures.

Conceptual Understanding

The following presents the results concept science notebook scores for the language status (i.e., ELL, former ELL, and non-ELL) and gender groups for each domain (i.e., Domain 1 = physical science; Domain 2 = earth/space science; Domain 3 = life science) separately. Descriptive statistics, homogeneity of variance results, and ANOVA summary tables are presented for each domain.

Domain 1: Physical Science

Descriptive statistics of concept science notebook scores for the language status and gender groups for Domain 1 are listed in Table 11.

Table 11

Descriptive Statistics of Language Status Groups and Gender (Domain 1)

		N	Mean	Std. Deviation	Skewness		Kurtosis	
		Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ELL								
	Female	4	2.3675	.35528				
	Male	3	1.7300	.28054				
	Total	7	2.0943	.45328	.241	.794	-.175	1.587
Former ELL								
	Female	5	2.8060	.35725				
	Male	4	2.2500	.39488				
	Total	9	2.5589	.45625	.000	.717	-1.843	1.400
Non-ELL								
	Female	5	2.5693	.32925				
	Male	5	2.2133	.4461				
	Total	10	2.4840	.24789	-.700	.687	.250	1.334
Total Gender								
	Female	14	2.5693	.32925	.295	.597	-.327	1.154
	Male	12	2.2133	.44461	.073	.637	-.855	1.232
	Total	26	2.4050	.41964				

According to Table 11, in total, 7 ELL students, 8 former ELL students, and 10 non-ELL students had scores for the three time points; therefore the ANOVA was very nearly balanced, especially between the former ELL and non-ELL group. Table 11 also indicates the absolute values of skewness and kurtosis statistics are between +/- 1 and +/- 2 for both the language status groups and the total gender groups, acceptable values for psychometric purposes (Cutting, 2012), which underlie ANOVA assumptions (Thompson, 2008). This means the data are normally distributed for Domain 1.

The statistics also indicate that the ELL group had a lower overall mean (ELL Mean = 2.0943) than the former ELL group (Mean = 2.5589) and non-ELL (Mean = 2.4840) group, but the former ELL had a higher mean than the non-ELL group. There was a greater mean difference between genders for the ELL group (2.3675 - 1.7300 = .6375) than for the former ELL group (2.8060 - 2.2500 = .556) and a greater mean difference for the former ELL group than for the non-ELL group (2.5693 - 2.2133 = .356), with all means higher for females than for males within each group and in total (Female Total Mean = 2.5693, Male Total Mean = 2.2133).

Having established the data are normally distributed, a 3 x 2 ANOVA analysis was run for Domain 1 (Physical Science) to compare the dependent variable (i.e., conceptual understanding scores) across language status (ELL, former ELL, non-ELL) and gender groups. Levene's Test of Equality of Error Variances was not statistically significant (*p* = .106), meaning the data met the homogeneity of variance assumption. Table 12 presents the 3 x 2 ANOVA summary table for Domain 1.

Table 12

Two-Way ANOVA Results Comparing Language Status and Gender Groups for Conceptual Scores in Domain 1

Source	SOS	df	Mean Square	F	Sig.	Partial Eta Squared
Group	1.055	2	.527	5.105	.016	.338
Gender	1.027	1	1.027	9.944	.005	.332
Group * Gender	.506	2	.253	2.448	.122	.197
Error	2.066	20	.103			
Total	4.654	26				

Note: R Squared = .531 (Adjusted R Squared = .413)

As can be noted in Table 12, the interaction effect of group * gender was not statistically significant (*p* = .122). Therefore, main effects of group and gender were analyzed. Main effects for group (*p* = .016; partial eta squared = .338) and gender (*p* = .005; partial eta squared = .332) were significant and substantial[6].

A Tukey post-hoc test, appropriate for comparisons interested in simple contrasts and with smaller sample sizes in order to maintain power (Thompson, 2008), was run to determine where the differences existed between the language status groups. Mean differences between the ELL and former ELL group (*p* = .025; std. error = .16199) were statistically significant. Mean differences between the ELL and non-ELL group were not statistically significant (*p* = .058; std. error = .15841) but close to being statistically

---

[6] Partial eta squared, or the effect size magnitude is being gaged according to the commonly used guidelines proposed by Cohen (1988, pp. 284-287): .01=small effect, .06=moderate effect, .14=large effect.

significant. Mean differences between the former ELL and non-ELL group were not statistically significant (*p* = .869; std. error = .14769).

Figure 3 illustrates differences between the gender groups by language status group.

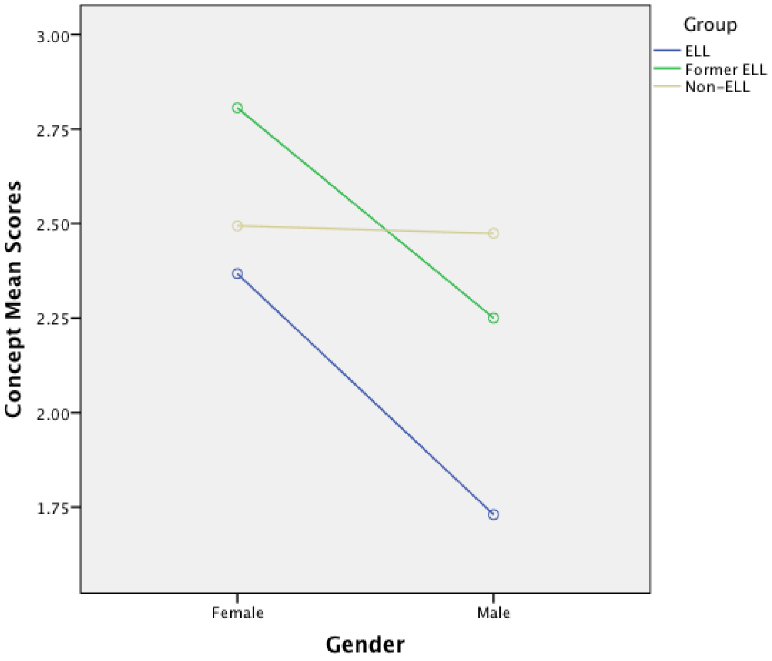

Figure 3. Domain 1 concept scores by group and gender.

Figure 3 illustrates the differences in means noted in the descriptive statistics, with greater mean differences between genders for the ELL group than for the former ELL group and a greater mean difference for the former ELL group than for the non-ELL group. In fact, the gender difference between the non-ELL group is much smaller than the difference between the ELL and former-ELL group. Female means were, overall, higher than male means.

Domain 2: Earth/Space Science

Descriptive statistics of concept science notebook scores for the language status (ELL, former ELL, non-ELL) and gender groups for Domain 2 are listed in Table 13.

Table 13

Descriptive Statistics of Language Status Groups and Gender (Domain 2)

		N	Mean	Std. Deviation	Skewness		Kurtosis	
		Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ELL								
	Female	4	2.5200	.63156				
	Male	3	1.5667	.45092				
	Total	7	2.1114	.72587	.442	.794	.280	1.587
Former ELL								
	Female	5	2.5780	.25489				
	Male	4	2.4000	.43734				
	Total	9	2.4989	.33617	-.184	.717	-1.175	1.400
Non-ELL								
	Female	5	2.7060	.40679				
	Male	5	2.9900	.35812				
	Total	10	2.8480	.39109	-.340	.687	-.185	1.334
Total Gender								
	Female	14	2.6071	.41155	-.215	.597	-.070	1.154
	Male	12	2.4375	.69416	-.413	.637	-.356	1.232
	Total	26	2.5288	.55455				

According to Table 13, in total, 7 ELL students, 8 former ELL students, and 10 non-ELL students had scores for the three time points; therefore the ANOVA was very nearly balanced, especially between the former ELL and non-ELL group. Table X4 indicates the absolute values of skewness and kurtosis statistics are between +/- 1 and +/-

2 for both the language status groups and the total gender groups, acceptable values for psychometric purposes (Cutting, 2012), which underlie ANOVA assumptions (Thompson, 2008). This means the data are normally distributed for Domain 2.

The statistics also indicate that the ELL group had a lower overall mean (Mean = 2.1114) than the former ELL group (Mean = 2.4989) and the former ELL group have a lower mean than the non-ELL group (Mean = 2.8480). Once again, there was a greater mean difference between genders for the ELL group (2.5200 – 1.5667 = .9533) than for the former ELL group (2.578 – 2.4000 = .178) and a greater mean difference for the former ELL group than for the non-ELL group (2.9900 - 2.7060 = .284), with means higher for females than for males for the ELL and former ELL group, but not for the non-ELL group in which the males had higher means. In total, female means are higher than male means (Female Total Mean = 2.6071, Male Total Mean = 2.4375). Yet, it is notable that males in the Non-ELL group scored highest and males in the ELL group scored lowest overall.

Having established the data are normally distributed, a 3 x 2 ANOVA analysis was run for Domain 2 (Earth/Space Science) to compare the dependent variable (i.e., conceptual understanding scores) across language status (ELL, former ELL, non-ELL) and gender groups. Levene's Test of Equality of Error Variances was not statistically significant (p = .713), meaning the data met the homogeneity of variance assumption. Table 14 presents the 3 x 2 ANOVA summary table for Domain 2.

Table 14

Two-Way ANOVA Results Comparing Language Status and Gender Groups for Conceptual Scores in Domain 2

Source	SOS	df	Mean Square	F	Sig.	Partial Eta Squared
Group	2.640	2	1.320	7.308	.004	.422
Gender	.501	1	.501	2.774	.111	.122
Group * Gender	1.559	2	.780	4.318	.028	.302
Error	3.612	20	.181			
Total		26				

As can be noted in Table 14, the main effect for group was statistically significant (p = .004) with a Tukey post-hoc test noting a statistically significant difference between the ELL and non-ELL groups (*p* = .006). Furthermore, the interaction effect of group * gender was statistically significant (*p* = .028).

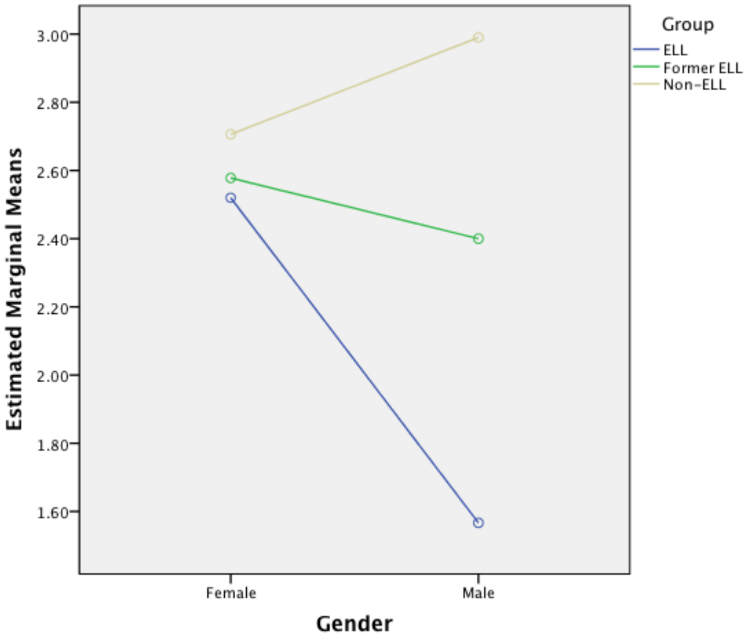

Figure 4. Domain 2 concept scores by group and gender.

Figure 4 data suggests a much greater mean difference among male students across the language status groups than among female students across the language status groups. In addition, females overall had higher means than males (though males slightly outperformed females in the non-ELL group).

Domain 3: Life Science

Descriptive statistics of concept science notebook scores for the language status (ELL, former ELL, and non-ELL) and gender groups for Domain 2 are listed in Table 15.

Table 15

Descriptive Statistics of Language Status Groups and Gender (Domain 3)

		N	Mean	Std. Deviation	Skewness		Kurtosis	
		Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
ELL								
	Female	4	2.7700	.45453				
	Male	3	2.6900	.78886				
	Total	7	2.7357	.55907	.821	.794	-1.055	1.587
Former ELL								
	Female	5	3.0700	.67786				
	Male	4	2.7875	.44709				
	Total	9	2.9444	.57173	.755	.717	-1.143	1.400
Non-ELL								
	Female	5	3.0140	.29194				
	Male	5	2.6540	.47705				
	Total	10	2.8340	.41836	-.683	.687	.444	1.334
Total Gender								
	Female	14	2.9643	.48182	.634	.597	-.753	1.154
	Male	12	2.7075	.50411	.456	.637	-.829	1.232
	Total	26	2.8458	.49958				

Concept scores for Domain 3 had the same number of students in the language status groups (i.e., ELL N = 7; former ELL N =8; non-ELL N = 10) and total gender groups (i.e., Female N = 14; Male N = 12) as Domains 1 and 2, indicating a nearly balanced design. Table 15 indicates the absolute values of skewness and kurtosis statistics are between +/- 1 and +/-2 for both the language status groups and the total gender groups, acceptable values for psychometric purposes (Cutting, 2012), which underlie ANOVA assumptions (Thompson, 2008). This means the data are normally distributed for Domain 3.

Like Domain 1, descriptive statistics for Domain 3 indicate that the ELL group had a lower overall mean (ELL Mean = 2.7357) than the former ELL group (Mean = 2.9444) and non-ELL (Mean = 2.8340) group, but the former ELL had a higher mean than the non-ELL group.

In Domain 3, there was a greater mean difference between concept scores between genders for the non-ELL group (3.0140 – 2.6540 = .36) than for the former ELL group (3.0700 – 2.7875 = .2825) and a greater mean difference for the former ELL group than for the ELL group (2.7700 – 2.6900 = .08). Domain 3, therefore, is the only domain in which ELLs have the least mean difference between concept scores across genders. Consistent with findings in Domain 1 and 2, however, overall mean concept scores were higher for females than for males (Female Total Mean = 2.9643, Male Total Mean = 2.7075).

Having established the data are normally distributed, a 3 x 2 ANOVA analysis was run for Domain 3 (Life Science) to compare the dependent variable (i.e., conceptual understanding scores) across language status (ELL, former ELL, non-ELL) and gender groups. Levene's Test of Equality of Error Variances was not statistically significant (*p* = .135), meaning the data met the homogeneity of variance assumption. Table 16 presents the 3 x 2 ANOVA summary table for Domain 3.

Table 16

Two-Way ANOVA Results Comparing Language Status and Gender Groups for Conceptual Scores in Domain 3

Source	SOS	df	Mean Square	F	Sig.	Partial Eta Squared
Group	.153	2	.077	.762	.762	.027
Gender	.364	1	.364	.266	.266	.062
Group * Gender	.082	2	.041	.864	.864	.014
Error	5.553	20	.181			
Total		26				

As can be noted in Table 16, the interaction effect of group * gender was not statistically significant (*p* = .864). Therefore, main effects of group and gender were analyzed. Main effects for group (*p* = .726) and gender (*p* = .226) were not significant, indicating that by Domain 3 at the end of the year, group and gender differences in conceptual understanding scores had disappeared.

Relationship between Language and Concept

As was noted in establishing the construct validity of the rubrics, correlations between language and concepts scores were positive, large, and significant for each domain (i.e., Domain 1 Pearson *r* = .843; Domain 2 Pearson *r* = .878; Domain 3 Pearson *r* = .810; all significant at *p* < .01). The following presents the results of Pearson *r* correlations between student language status (ELL, former ELL, non-ELL) and gender groups within each domain. Domains were analyzed separately in order to compare possible differences in correlations between domains, which imply both different time periods (i.e., beginning, middle, end of year) and different concepts (i.e., physical science, earth/space science, and life science).

Language Status Groups

Table 17 illustrates the results of Pearson *r* correlations between language and concept scores for each language status group within each domain.

Table 17

Correlations between Language and Concept Mean Scores for Language Status Groups within Each Domain

	Domain 1	Domain 2	Domain 3
Language status group	Pearson *r*	Pearson *r*	Pearson *r*
ELL	.816* (n = 7)	.796* (n = 7)	.945** (n = 7)
Former ELL	.896** (n = 9)	.851** (n = 9)	.716* (n = 9)
Non-ELL	.862** (n = 10)	.895** (n = 10)	.780** (n = 10)

Note: Domain 1 = Physical Science, Domain 2 = Earth/Space Science, Domain 3 = Life Science
* Correlation is significant at the .05 level (2-tailed).
** Correlation is significant at the .01 level (2-tailed).

According to Table 17, correlations between language and concept were positive and significant for each of the three language status groups within each of the domains. The ELL groups in Domain 3 had the largest correlation (Pearson *r* = .945; *p* = .005).

Gender

Table 18 illustrates Pearson *r* correlations between language and concept scores for gender groups within each domain.

Table 18

Correlations between Language and Concept Mean Scores for Gender Groups within Each Domain

	Domain 1	Domain 2	Domain 3
Gender Group	Pearson *r*	Pearson *r*	Pearson *r*
Female	.567* (n = 14)	.593* (n = 14)	.782* (n = 14)
Male	.925** (n = 12)	.568 (n = 12)	.879** (n = 12)

Note: Domain 1 = Physical Science, Domain 2 = Earth/Space Science, Domain 3 = Life Science; Significance for Males in Domain 2 is at the .054 level.
* Correlation is significant at the .05 level (2-tailed).
** Correlation is significant at the .01 level (2-tailed).

According to Table 18, correlations between language and concept were positive and significant for gender groups in all domains except for males in Domain 2 (Pearson *r* = .568; *p* = .054), though the significance was very close at the .054 level. At the same time, raw data indicates males in Domain 2 had overall low scores (Total Male Mean = 2.4375) and high standard deviations (Total Male SD = .69416) compared to females

(See Table 13), thus adding to the lack of significance of this particular value. The largest significant correlations were found for males in Domains 1 (Pearson $r = .925$; $p < .01$) and 3 (Pearson $r = .879$; $p < .01$). Females had smaller correlations than males in Domains 1 and 3.

Summary

The purpose of the present study was to (a) investigate the impact of science notebook writing on students' academic language development over time, across student language status and gender groups. Three individual time points (beginning/middle/end of science notebook kept over the course of one academic year) were rated; therefore, three time points of data were analyzed; (b) compare students' conceptual understanding across student language status and gender groups, and (c) explore the extent of relationship between students' academic language and conceptual understanding scores. With a total of two sets of 26 average mean scores, one for language and one for concept, data analysis in this chapter were reported in the following order: (a) descriptive statistics and normality check; (b) homogeneity of variance and sphericity (if appropriate) check; (c) interaction effect check and analysis if appropriate; (d) main effect analysis if appropriate; (e) post-hoc tests if appropriate; (f) tables and figures to guide analysis where needed; (g) results of Pearson *r* correlations for the final analysis. The following chapter will present discussions, limitations, recommendations, and conclusions.

CHAPTER V

DISCUSSION, LIMITATIONS, CONCLUSIONS AND RECOMMENDATIONS

Students' language proficiency parallels their ability to build upon science understanding as well as to demonstrate it (Kieffer et al., 2009; Lee & Lyukx, 2005). Researchers, therefore, have advocated integrating science and literacy instruction in the content-area classroom (Fang, 2006; Janzen, 2008, Lee, 2005). Moreover, Kieffer et al. (2009), in their seminal study regarding the academic achievement of ELL students in science, imply a strong parallel between diverse students' science achievement and their academic language ability. Few researchers explore the impact of literacy-integrated activities, such as science writing, on students' scientific conceptual understanding *and* language development scores (e.g., Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010). Even fewer researchers consider the role of science and literacy/writing integration on ELL and low-SES students' conceptual understanding and academic language development (e.g., Lee et al., 2009).

My study tracked the language development and conceptual understanding achievement (as reflected in science notebooks) of 26 fifth-grade students who received a science inquiry and literacy-integrated intervention. The students were part of a randomized, longitudinal, field-based NSF funded research project (NSF Award No. DRL - 0822343) that targeted ELL and non-ELL diverse students from low-SES backgrounds in a large urban school district in Southeast Texas. The results of my study can inform future researchers, educators, and policy-makers on the impact of literacy integrations, such as science notebooks, on the academic language development and conceptual understanding for populations of linguistically diverse and low-SES students.

Discussion

A comprehensive discussion of the result findings, linked to the previous literature, follows. The discussion is broken down by the study's research questions.

Research Question #1

Research question #1 was as follows: Over the course of 1 year, did ELLs, former ELLs, and non-ELL/low-SES 5th grade students make significant gains in academic language, and to what extent does the level of academic language across student language status (ELL, former ELL, non-ELL) and gender groups differ? The following discusses the findings for research question #1.

Time. ELLs, former ELLs, and non-ELLs did make significant gains in academic language over the course of one year as measured by their science notebook language mean scores. The main effect for time for all groups was statistically significant ($p = .000$) and substantial (partial eta squared = .604). Descriptive data also shows that all groups except non-ELLs increased in language mean scores between each of the three time points (i.e., Domain 1, Domain 2, and Domain 3). Non-ELL means dipped down from Domain 2, the second period ($M = 2.9120$), to Domain 3, the third time period ($M = 2.8730$), but finished with higher scores in Domain 3 ($M = 2.8730$), than in Domain 1 ($M = 2.6310$). ELLs, who traditionally struggle with academic language and lag behind their English speaking peers (Kieffer et al., 2009; Lee & Lyukx, 2005), demonstrated an increase in mean scores for each of the time periods, with ELLs catching up to the non-ELL. Moreover, former-ELLs exceeded the means of the ELL and non-ELL group.

Researchers (Amaral et al., 2002; Lee et al., 2005; Lee et al., 2009) who examine the academic language development of students receiving science inquiry and literacy-integrated interventions with samples of ELL, former ELL, and non-ELL interventions, support the above findings since they all reported gains in students' academic language development over time. For example, Lee et al. (2005) reported literacy gains over the course of one year, measured by pre-post test writing scores for all 4th grade students participating in their science inquiry and literacy integration, as being statistically significant and substantial for writing form ($p = .001$; Cohen's $d = 1.11$). Similarly, Lee et al. (2009) reported a statistically significant mean gain writing form score of .84 ($p < .001$) over the course of one year for all 3rd grade students in their science and literacy integration. Last, Amaral et al. (2002) reported descriptive statistics percentages showing an increase in passing rate scores measured by a standardized writing proficiency test for all students each year they participated in their science and literacy-based integration (e.g., ELL year 0 = 52.5% passing, ELL year 4 = 78.3% passing; former ELL and non-ELL year 0 = 64.4% passing, former ELL and non-ELL year 0 = 89.7% passing).

Language status group differences. Overall, language status groups differed between ELL and non-ELL groups, but not between ELL and former ELL or between former ELL and non-ELL groups. The finding regarding differences between ELL and non-ELL groups are supported by Lee et al. (2005) who reported statistically significant differences in the post-test writing form scores between ELL ($t = - 4.76$, $p < .001$) and non-ELL students but no statistically significant differences between the former ELL and non-ELL students (t = - 2.84, *p* < .001).

The findings regarding non-significance between former ELLs and ELLs, on the other hand, differed from previous studies (Lee et al., 2008; Lee et al., 2009), though caution is warranted in comparing the present study to previous studies. For example, in Lee et al.'s (2009) study the researchers noted former ELLs, "tend to perform higher than [ELLs]" (p. 160) and therefore included former ELLs with the sample of non-ELLs for data analysis. Consequently, the researchers were not able to measure possible statistically significant differences in academic language scores between the former ELL group and ELL and the non-ELL groups. The present study, on the other hand, was able to consider each group (i.e., ELL, former ELL, and non-ELL) separately, adding a new dimension to research in the area of science and literacy integration with respect to the former-ELL group. That is, according to the findings of the present study, former-ELLs appear to lie somewhere in the middle in terms of academic language development between ELL and non-ELL groups.

Gender differences. Differences in academic language mean scores for females and males were statistically significant (*p* = .041, partial eta squared = .193). Females had overall higher average mean academic language scores than male students at time points 1 and 2, but male and female academic language scores evened-out at time point 3. These findings differ from Lee et al. (2005) who reported no gender difference in academic language in pre-posttest writing measures for their sample of ELL and non-ELL 3rd and 4th graders. To a certain extent, the findings in the present study are exploratory. Comparison of measurement findings from one previous study (e.g., no gender difference) and the present study (i.e., a female advantage) should be taken with caution given that no previous researchers looking at students' academic language based on *science notebook* scores (vs. pre-post tests) consider gender as a variable (e.g., Ruiz-Primo et al., 2004).

Past studies with specific writing in science interventions still provide insight into the results of the present study. Patterson (2001), for example, found girls preferred more rigid planning structures for science writing than boys did. The intervention in the present study mirrors more rigid and structured writing formats. It could be that female scores were higher within the construct of academic language because of the structured writing format of the intervention used in the present study.

The present study, situated in the context of science education, presents noteworthy findings with respect to language learning favoring female students but with male students catching up to their female counterparts by the last time point. Most importantly, both groups demonstrated growth in academic language achievement over time over the course of the academic year in the science inquiry and literacy integrated intervention. In the end (i.e., time point 3), the initial achievement gaps in academic language L2 learning for gender groups disappeared – a sign that the intervention had a positive effect on student learning, regardless of gender.

Research Question #2

Research question #2 was as follows: To what extent does the level of conceptual understanding across student language status (ELL, former ELL, non-ELL) and gender groups for science notebook entries differ? The following discusses the findings for research question #2.

Note on conceptual understanding. Conceptual understanding, as a construct, was analyzed separately for each domain, given that some domains may be more difficult to grasp than others (vs. tracking growth over time as was done for academic language as a construct) and given that the units of analysis included unique science notebook entries vs. a pre post-test design where the same writing prompt is given to the students at the beginning and then end of the year. The following will discuss the findings with respect to students' conceptual scores according to each domain and will then provide of summary of all the findings for students' conceptual scores.

Domain 1: Physical science. The following discusses findings from Domain 1 with respect to conceptual understanding. Findings regarding language status and gender group are discussed separately since not interaction effects were detected between the variables.

Language status groups. In Domain 1 (Physical Science) mean differences between the ELL and former ELL group and between the ELL and non-ELL group were statistically, or nearly statistically, significant. Mean differences between the former ELL and non-ELL group, however, were not statistically significant.

Lynch et al.'s (2005) work most closely supports the present study's findings. The researchers' ANCOVA analysis noted ELL students performed significantly lower than non-ELL and former ELL students on conceptual science achievement scores (but no significant differences were noted between non-ELL and former ELL groups). Lee et al.

(2005), on the other hand (the only other study which statistically breaks down comparisons of science concept scores for ELL, former ELL, and non-ELL groups), reported that ELL and former-ELL students scored significantly lower than non-ELL students at posttest (i.e., there were significant differences between former ELL and non-ELL groups unlike the present study's findings).

Of course, comparisons between the present study and previous studies warrants caution given that Lynch et al. (2005) and Lee et al. (2005) considered the science conceptual development of ELL, former-ELL, and non-ELL students *over time*. The present study looked at conceptual achievement at separate time points and specific science domains (vs. overall science knowledge from pre and posttests) based on science notebook scores (vs. standardized achievement tests).

Gender groups. In Domain 1, gender differences were significant and substantial (p = .005; partial eta squared = .332) with females having a higher overall concept mean scores (M = 2.5693) than males (M = 2.2133). Lee et al.'s (2005) and Lynch et al.'s (2005) studies, unlike the present study, reported no significant difference between gender groups for conceptual scores with samples of ELL students. Specifically, Lee et al. (2005) noted no significance difference in student growth rates according to gender from pre to posttest science achievement scores after an HLM analysis and Lynch et al. (2005) noted no significant difference in post-test scores between genders after an ANCOVA analysis.

In relating the present study's findings to past research, the following two points should be considered: (a) only two previous research studies with ELL samples in compared science concept scores *and* considered gender as a variable (b) the studies reported concept scores from standardized tests in pre-posttest designs, while the present study considers conceptual scores from science notebooks at the beginning of the year. It could be that gender differences in Lee et al.'s (2005) and Lynch et al.'s (2005) study, for example – should they have existed - disappear at the end of the year given a strong intervention, regardless of the difficulty of conceptual domains.

Domain 2: Earth/space science. The following discusses findings from Domain 2 with respect to conceptual understanding. The main effect of language status groups is discussed first, followed by a discussion of the gender and language status group interaction effect.

Language status group main effect. In Domain 2, Earth/Space Science, noted as one of the most difficult conceptual domains for students in science according to state achievement data (J. Jackson, Personal Communication, August 30, 2012), the main effect for language status group was statistically significant (p = .004) with a Tukey post-hoc test noting a significant difference between the ELL and non-ELL groups (*p* = .006). As has been noted in the discussion regarding Domain 1, Lee et al.'s (2005) and Lynch et al.'s (2005) work supports this finding of *overall* significant differences in concept scores for groups of ELL vs. non-ELL students. In addition, Lee et al. (2008), who included former ELL and non-ELL students in the same category for their HLM analysis, also supports the findings of the present study. The researchers reported a coefficient for the treatment ELL students which differed significantly from zero (-1.64) meaning that students in the former ELL/non-ELL group scored significantly higher than students who were classified ELL.

Gender and language status group interaction effect. In Domain 2, the interaction effect of group * gender was statistically significant ($p = .028$). Analysis of the data suggests male students varied much more across language status groups than female students. Females, also, had overall higher means. Previous work in this field has not considered possible interactions between language status group and gender. For example, Lynch et al. (2005) used ANCOVA to consider interactions between their "curriculum conditions" (p. 929) and other variables (e.g., gender, language status group) but not specifically between gender and language status groups. It is not clear, therefore, how the present finding fits into previous research; analysis of the relationship between language and concept by domain and gender might inform this finding (see discussion section to research question #3).

Domain 3: Life science. The following discusses findings from Domain 3 with respect to conceptual understanding. Findings regarding language status and gender group are discussed separately since no interaction effects were detected between the variables.

Language status groups. In Domain 3 (Life Science) the main effect for language status groups ($p = .864$) was not statistically significant. The findings differ from previous researchers (e.g., Lee et al., 2005; Lee et al., 2008; Lynch et al., 2005) who, as has bee discussed in detail under the discussion of Domains 1 and 2, noted significant differences in conceptual science scores between language status groups at the end of their intervention studies (even if the growth rates for groups are similar). Once again, comparisons between the present study and previous studies should be made with caution since previous studies looked at pre-posttest scores at the end of an intervention while the present study looks at conceptual scores for specific domains at specific times of the year.

Even though the construct of "concept" was not measured over time in the present study, the fact that the language status groups did not differ late in the school year (i.e., Domain 3) may indicate positive effects from the intervention on students' achievement – a finding quite different from that of previous work where ELLs score significantly below the non-ELL and former ELL groups, even at the end of the school year (e.g., Lee et al., 2005; Lee et al., 2008; Lynch et al., 2005).

Gender groups. In Domain 3 (Life Science), the main effect for gender groups (*p* = .226) was not statistically significant. In this case (unlike Domains 1 and 2) findings for conceptual scores align with previous researchers' (Lee et al., 2005; Lynch et al., 2005) who reported no significant difference between gender groups for conceptual scores with samples of ELL students. Again, in the context of this study, even though student conceptual academic achievement was not measured over time, it is worth noting that gender differences existed in Domain 1 and 2 but disappeared by Domain 3. Gender differences may have become insignificant at the end of the year due to the intervention, though conclusions cannot be made on this point given the design of the present study (i.e., the present study cannot measure conceptual growth over time).

Research question #2 summary. The following summarizes the findings from the three science domains. The summary allows for a comprehensive view of results in the realm of student conceptual understanding.

Language status groups. In summary, ELLs scored significantly lower than non-ELLs in Domains 1 and 2, aligning with previous research findings on science concept scores between these two language status groups (Lee et al., 2009; Lee et al., 2008; Lynch et al., 2005). Former ELLs scored significantly higher than ELLs in Domain 1, aligning with Lynch et al.'s (2005) findings, differing with Lee et al.'s (2005) findings, but contributing to the small body of research considering former ELLs. No statistically significant differences were found among the language status groups (ELL, former ELL, and non-ELL) in Domain 3, which differs from previous researchers' findings (Lee et al., 2005; Lee et al., 2009; Lee et al., 2008; Lynch et al., 2005).

Gender groups. Females scored statistically significantly higher than males in Domain 1 and had higher overall means in Domain 2, unlike previous research with samples of ELL students, which note no differences in science concept scores between genders (Lee et al., 2005; Lynch et al., 2005). Researchers have noted an overall female advantage in both first and second language ability (e.g., Brantmeier, Schueller, Wilde, & Kinginger, 2007), so it is possible that the nature of the assessment medium – that is, a medium highly dependent on language/writing – affected the results, giving females the advantage in Domains 1 and 2. In Domain 3, however, no significant differences were found between the gender groups, aligning to previous researcher's findings (Lee et al., 2005; Lynch et al., 2005).

Note. In comparing previous research to the present study, caution in result comparisons is warranted due to the experimental nature of the present study. Previous researchers (e.g., Lee et al., 2009; Lee et al., 2008; Lynch et al., 2005) use pre-post test scores from standardized achievement tests at the beginning and end of the year to track and compare students' conceptual growth over time. The present study uses a rubric to rate student science notebook entries representing various conceptual domains, thus not being able to measure students' conceptual growth over time. Previous research, nonetheless, informs the experimental results of the present study. Arguably, the present study adds to the limited work in this area of conceptual development comparisons of ELL, former ELL, non-ELL, and gender groups in the context of science learning. Of particular interest is the fact that, in Domain 3, language status group differences and gender differences disappeared.

Research Question #3

Research question #3 was as follows: To what extent are academic language and conceptual understanding related as reflected in science notebook entry scores, and how do the relationships compare between student language status (ELL, former ELL, non-ELL) and gender groups? The following discusses the findings for research question #3.

Academic language and conceptual understanding. Correlations between language and concepts scores were positive, large, and significant for each domain. Overall, these findings align with theorists (e.g., Halldén, 1999) and researchers (e.g., Kieffer et al., 2009) who note connections between academic language and conceptual understanding. Specifically, correlations were strongest for Domain 2 (Earth/Space Science), which had the least amount of writing/language, and weakest for Domain 3 (Life Science), which had the most amount of writing/language.

Correlations findings by domain are supported by Ruiz-Primo et al. (2004)'s work. Ruiz-Primo et al. (2004) noted larger correlations for language and concept between notebook entries they considered to reflect procedural understanding (e.g., reporting graphs and tables, which require a more structured format and less language) vs. notebook entries they considered to reflect more conceptual understanding (e.g., definitions and explanations which generally follow a less structured format and include more language). In the present study, notebook entries in Domain 2 (Earth/Space Science) – though supposed to be more conceptual, or perhaps because they were more difficult conceptually – included highly structured entries with little language. Students pasted the same cut outs of the moon or sun and added short labels and occasional observations and explanations (many of which included the same wording, as students seemed to be copying explanations off of the board). Thus, the entries' format and execution reflected more procedural understanding than conceptual understanding. On the other hand, notebook entries in Domain 3 (Life Science) consisted of less structured and more prolific written observations and explanations regarding the topics the students were learning about (consequently, the language and content varied more from student to student). Domain 3 therefore typified the most conceptual understanding type of entry in the sense of having the most "free" and prolific use of language. It stands to reason, in line with Ruiz-Primo et al.'s (2004) findings, that entries in Domain 2, which appear to reflect more procedural execution/understanding, should have higher language and concept correlations than entries in Domain 3, which appear to reflect more conceptual execution/understanding do.

Ruiz-Primo (2004) reasoned that, “Entries that focused more on process skills require better communication skills than those focusing on definitions or examples” (p. 1495). It is not clear what the researchers meant by “better communication skills,” though it is possible they mean more “precise” communication skills in terms of format, structure, and academic word usage. One could argue the opposite – that is, entries focused more on conceptual skills require better communication skills. Key’s (2000) qualitative work would likely agree with this statement. However, Ruiz-Primo et al.’s (2004) and the present study’s rubric calls for quantitative measurements, which currently distinguish between content and language (including grammar and form in the case of the present study’s rubric adaptation). Therefore, it is more plausible to state that entries that focused on process skills require less language than those focusing on concept skills, at least, when measured by a science notebook rubric. Furthermore, the more language the student was using in an entry, the more likely the correlation between language and concept would be low. More language use in an entry appears to begin to tap into dimensions of language that may not correlate with conceptual understanding as measured by a science notebook rubric.

Language status groups. Correlations between language and concept were all positive and significant for each of the three language status groups within each of the domains. The largest correlation was found in the ELL group in Domain 3. The latter finding may due to the following two observations (a) ELL language scores significantly increased by the end of the year (as noted in the previous findings of this study), so whatever language ELL students at this point in time were using was to their advantage

(b) From observation, ELL notebook entries, overall, used less language than non-ELL and former ELL notebook entries. Therefore, if the idea previously presented that less language use leads to higher correlations between language and concept, it would make sense that the highest correlation between language and concept would be found in entries which use less language (i.e., the ELL notebook entries).

Gender groups. Correlations between language and concept were positive and significant for gender groups in all domains except for males in Domain 2, though the significance was very close at the .054 level. It is possible that the correlation was not significant because raw data show overall low scores among male students and high standard deviations in Domain 2. Females had smaller correlations than males in Domains 1 and 3 and the largest significant correlations were found for males in Domains 1 and 3. Given the previous theory that the use of more language in an entry lowers the language and concept correlation, it follows that: If females use more language than males, the correlations between language and concept should be lower for females than for males. Last, the findings provide some insight the gender interaction effect in Domain 2, where females had overall higher concept mean scores (and higher language mean scores) than males. It is possible that females, even if exhibiting more language use in the entries, had higher correlations simply because male students varied much more across language status groups. The variation, therefore, does not allow a clear picture of the dynamics of language and concept. Having a larger sample might help clarify findings.

Other Findings

In addition to the above, there are other findings worth noting in this study. First, a marginally significant interaction effect was noted for time and language status group with ELLs apparently making the most growth in language scores, followed by former ELLs, and then by non-ELLs. Even though the present study cannot compare the intervention group to the control group, it is nonetheless noteworthy that students exposed to an intervention so closely targeting the academic English language growth of students should have the greatest impact on ELL students. So, while the intervention seemed to help *all* students (e.g., former ELL and non-ELL as well), the intervention had the greatest impact on the ELL students.

Second, raw data showed an interesting pattern of mean differences in conceptual scores between genders across language status groups. In Domains 1 and 2, mean differences between genders were greatest for the ELL group, followed by the former ELL group, and then the non-ELL group (i.e., the non-ELL group had the least mean differences between genders). Disparity between genders was therefore greater for ELL students than former ELL and non-ELL students. In Domain 3, however, the results were mirrored. That is, mean differences between genders were greatest for the non-ELL group, followed by the former ELL group, and then by the ELL group. It appears then, that the intervention had some effect in evening out the gender differences for ELLs but the opposite effect for the non-ELL group. Because the sample size was so small, however, it was not possible to see if the disparities were significant.

Limitations

As has been noted, the present study includes a sample of fifth grade students from low-income families, some of which were ELLs, in a single community. Therefore, the results are not generalizable given that the sample may not represent the population of all fifth grade, low-SES and/or ELL students. However, generalizability may be inferred to students with similar characteristics to those in this study. At the same time, science notebook writing samples form the control classrooms were not collected for the present study because the control classrooms were not expected to implement science notebook writing in science (even though the larger study included an intervention and control group). Therefore, it is not possible to conclude whether student results in language achievement and conceptual understanding are a result of the intervention. Still, the analysis and results of the present study provide critical insight into the role science notebook writing played on students' academic language development and conceptual understanding within a science inquiry intervention.

Conclusions and Recommendations

The following discusses conclusions derived from the findings of this study. It also provides recommendations for future research and practice.

Science Notebook Rubric

The science notebook rubric used in this study indicates high validity and reliability for rating science notebook samples from a population, which includes ELL and low SES students. These findings show how, when created/modified thoroughly with strong validity (i.e., content validity and construct validity) and reliability measures (i.e., G theory), this particular rubric can yield valid and reliable measures on academic language and conceptual scores. Moreover, the findings add to the work of Ruiz-Primo et al. (2004) who created the original rubric and proved it could be applied reliably in their own research studies. Future work could consider other forms of construct validity, such as concurrent validity measures in which science notebook language and concept scores are compared to language and content scores from standardized tests. This kind of work would allow for comparisons in scores between assessments that are more *proximal*, or close to the specific classroom curriculum (e.g., teacher created tests, science notebook entries), and more *distal*, or further from the classroom curriculum (e.g., state and national standardized tests) – comparisons, as noted by Ruiz-Primo et al. (2004), critical to inform whether science interventions is having an impact on student achievement at all levels of assessment (i.e., proximal and distal).

The science notebook rubric in this study can be used/modified for future research studies with populations of ELL and low-SES students. In this way, studies could add to research on science inquiry and literacy interventions with ELL and low-SES students (e.g., Lee et al., 2005; Lee et al., 2009) and allow researchers to further understand the role of writing in the academic language development and conceptual understanding of ELL and low-SES students. Future studies could also consider the practicality of the instrument by training classroom teachers to use the instrument and then calculating reliability estimates from within classroom use. The instrument could inform both research and practice as an instrument for both summative and formative assessment in the classroom. This kind of analysis would prove particularly useful in science interventions like the one used in MSSELL which, like Akkus, et al.'s (2007) Science Writing Heuristic, use a structured lesson plan template and scaffold meant to guide students' conceptual understanding through a lesson sequence. The benefits of such a process on ELL and low-SES students science language development could also be tracked by using the present study's rubric.

Writing in Science and Academic Language

As measured by a science notebook rubric, writing in science does help increase students' academic language over time. This finding is in line with past studies considering the impact of writing in science on students' academic language with samples that include non-ELL students (Patterson, 2001; Ruiz-Primo et al., 2004) and ELL students (Amaral et al., 2002; Lee et al., 2005; Lee et al., 2009). Clearly, it is advantageous to include literacy activities in content-area classrooms in order to promote the academic language development of students. Future studies should consider mixed-method designs in which both quantitative and qualitative measure of academic language could be conducted to provide a more holistic picture of the impact of science notebook writing on students' academic language development.

In line with previous research, this study noted significant differences in academic language scores between ELL and non-ELL students (Amaral et al., 2002; Lee et al., 2005; Lee et al., 2009; Lynch et al., 2005). However, more research on former ELLs is needed as a point of comparison between the findings of this study and other studies. Future studies should consider analyzing former ELLs separately from ELL and non-ELL students. Based on this study's findings, former ELLs, like ELLs and non-ELLs alike require scaffolding of academic language in the classroom.

In the writing medium of science notebooks, females in all language status groups seemed to have an advantage over males in terms of academic language. Still, the results of this study show the intervention helped close whatever gender gap existed at the beginning and middle of the year. Therefore, while females advantage in a literacy-based medium aligns with research on gender differences regarding literacy ability (Brantmeier et al., 2007), it is revealing that interventions such as this one evened-out the playing field for male students over time. Future work should consider tracking gender differences at various times points of the year (and in different domains) vs. just as pre- posttest as one past study with ELL samples has done (Lee et al., 2005) in order to inform research and practice on instruction that provides gender equity in the science classroom, regardless of the intervention.

Writing in Science and Conceptual Understanding

As measured by a science notebook rubric, writing in science appears to contribute to the conceptual understanding of students in different language status and gender groups. These findings are in line with past studies considering the impact of writing in science on students' conceptual understanding with samples which include non-ELL students (Akkus, et al., 2007; Gunel et al., 2009; Hand, Gunel, & Ulu, 2009; Mason, 2001; Mason & Boscollo, 2000; Keys et al., 1999; Patterson, 2001; Ritchie et al., 2011) and ELL students (Amaral et al., 2002; Fradd et al., 2001; Lee et al., 2005; Lee et al., 2008; Lynch et al., 2005). Certainly, conceptual understanding in science benefits from literacy-based activities such as science notebooks.

Like previous studies (Lee et al., 2009; Lee et al., 2008; Lynch et al., 2005), non-ELLs had an overall advantage over ELLs in terms of conceptual scores (except for Domain 3) in the present study. Unlike previous studies (Lee et al., 2005; Lynch et al., 2005), females had an overall advantage over males in terms of conceptual scores (except for Domain 3) in the present study. Therefore, though the present study could not track conceptual development over time, it is worth noting that conceptual scores differed significantly among language status groups and between gender groups in Domains 1 and 2 but not in Domain 3.

Given the above findings and conclusions, future studies would greatly benefit from thinking through how conceptual domains could be measured over time by means of a science notebook (e.g., considering overarching scientific concepts of “patterns” or “systems” across units). Future studies could also consider mixed-method study designs that combine using the rubric with observations of students’ writing over time. In specific, studies that focus on the writing processes and pedagogies with ELL and low-SES learners in the science classroom and its impact on student learning, in the context of the science notebook – a medium which can combine scaffolded creative and conventional forms of writing like in Lee and colleagues’ (e.g., Fradd et al., 2001; Lee et al., 2005; Lee et al., 2008) work – would prove beneficial. Within the studies, researchers should consider tracking gender differences overall and between language status groups in order to add to the sparse research in this area. Practitioners should note the benefits of integrating thoughtful literacy activities (which serve as scaffolds for student learning) in the content-area classrooms for ELL and low-SES students.

Relationship between Language and Concept

The relationship between language and concept is evident from the findings of this study, given overall large and significant correlation scores between language and concept scores. The findings verify theorists' (e.g., Halldén, 1999) and researchers' (e.g., Kieffer et al., 2009) findings. However, the relationship is complex. Domain 2, in particular, proved perplexing in its interaction effect between group and gender and the close, but non-statistically significant correlation between language and concept for the male students.

In an attempt to understand the pattern of correlations between language and concept, this study proposes the theory that writing entries that contain more language begin to tap into different aspects of language, resulting in lower correlations between language and concept scores. The theory is based the numerical results of this study, Ruiz-Primo et al.'s (2004) findings, and the observation that, in the present study, entries reflecting scripted work/less language resulted in higher correlations overall between language and concept than entries reflecting less scripted work/more language.

Future work is needed to test the theory. In specific, studies could qualitatively quantify the amount of writing in science notebook entries and compare them to each other and to language and concept correlations by domains, gender groups, and even language status groups. Furthermore, a larger sample of student science notebooks would allow comparisons to see if there are statistically significant differences between language and concept correlations between gender groups, among language status groups, and between gender groups within language status groups. Furthermore, these differences could be explored within the different science domains. Last, time could be invested in further refining the rubric so that it reflects the interconnectedness of language and concept in an attempt to better tap into the constructs regardless of the domain or entry type.

Concluding Remarks

The present study introduces an adapted instrument (Ruiz-Primo et al., 2004) for rating science notebook entries from samples that include ELL and low-SES students with potential for high reliability (i.e., with proper training for raters). The study adds to research that has yet to calibrate science notebook instruments with populations of ELL and low-SES students (e.g., Ruiz-Primo et al., 2004; Ruiz-Primo et al., 2010) – populations currently most at risk for academic failure in science (Kieffer et al., 2009; NCES 2011; TEA, 2011).

The present study also provides an analysis of the effectiveness of the MSSELL (NSF Award No. DRL - 0822343), science and literacy integrated curriculum based on science notebook scores from a population of ELL and low-SES students. The study adds critical insight to research on science inquiry and literacy interventions with ELL and low-SES students (e.g., Lee et al., 2005; Lee et al., 2009). The findings from this study allow researchers to further understand the role of writing in the academic language development and conceptual understanding of ELL and low-SES students –

namely, science notebook writing can increase diverse students' academic language and reflects students' conceptual understanding. Furthermore, academic language and conceptual understanding are related, though more work is needed to decipher subtle differences between the constructs, especially with respect to measurement instruments and methods. Regardless, researchers and practitioners can note how writing in science, implemented in an effective context (e.g., scaffolding; science inquiry) and with quality teaching (e.g., alignment to standards; structured/consistent instruction), yields effective learning outcomes in the realm of academic language and informs the conceptual understanding for populations of ELL and low-SES students.

REFERENCES

Abella, R. (2005). An examination of the validity of English-language achievement test scores in an English Language Learner population. *Bilingual Research Journal, 29*(1), 127-144.

Akkus, R., Gunel, M., & Hand, B. (2007). Comparing an inquiry-based approach known as the Science Writing Heuristic to traditional science teaching practices: Are there differences? *International Journal of Science Education, 29*(14), 1745-1765.

Amaral, O. M., Garrison, L., & Klentschy, M. (2002). Helping English learners increase achievement through inquiry-based science instruction. *Bilingual Research Journal, 26*(2), 213-239.

August, D., Branum-Martin, L., Cardenas-Hagan, E., & Francis, D. (2009). The impact of an instructional intervention on the science and language learning of middle grade English language learners. *Journal of Research on Educational Effectiveness, 2*(4), 345-376.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English Language Learners. *Learning Disabilities Research and Practice, 20*(1), 50-57.

Ausubel, D. P., Novak, J. S., & Hanesian, H. (1978). *Educational psychology: A cognitive view.* New York, NY: Hold, Rinehart & Winston.

Baddeley, A. D. (1976). *The psychology of memory.* New York, NY: Harper & Row.

Bacharach, V. R., Baumeister, A. A., & Furr, R. M. (2003). Racial and gender science achievement gaps in secondary education. *The Journal of Genetic Psychology, 164*(1), 115-126.

Berkenkotter C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power.* Hillsdale, NJ: Erlbaum.

Brantmeier, C., Schueller, J., Wilde, J. A., & Kinginger, C. (2007). Gender equity in foreign and second language learning. In S. S. Klein (Ed.), *Handbook for achieving gender equity through education (2nd edition).* (pp. 305-334). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, G., & Irby, B. J. (2000). *The principal portfolio.* Thousand Oaks, CA: Corwin Press.

Butler, M. B., & Nesbit, C. N. (2008). Using science notebooks to improve writing skills and conceptual understanding. *Science Activities, 44*(4), 137-145.

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Powell, C., Westbrook, A. Landes, N., Spiegel, S., McGarrigle Stuhlsatz, M., Ellis, A., Resch, B., Thomas, H., Bloom, M., Moran, R., Getty, S., & Knapp, N. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications executive summary.* Colorado Springs, CO: Biological Sciences Curriculum Study.

Calkins, L. (1994). *The art of teaching writing.* Portsmouth, NH: Heinemann.

Chen, X. (2009). *Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education* (NCES 2009-161). Washington, DC: National Center for Educational Statistics.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measurements.* New York, NY: Wiley.

Cloud, N., Genesee, F., & Hamayan, E. (2009). *Literacy instruction for English language learners: A teacher's guide to research-based practices*. Portsmouth, NH: Heinemann.

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

Collier, V. P., & Thomas, W. P. (1989). How quickly can immigrants become proficient in school English? *Journal of Educational Issues of Language Minority Students, 5*, 26-38.

Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology, 106(*6), 1691-1730.

Cummins, J. (1981). Empirical and theoretical underpinnings of bilingual education. *Journal of Education, 163*(1), 16-29.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*(1)*,* 18-36.

Cutting, J. C. (2012). *SPSS: Descriptive statistics*. Retrieved from http://psychology.illinoisstate.edu/jccutti/138web/spss/spss3.html

Daniels, H., Zemelman, S., & Steineke, N. (2007). Content-area writing: Every teacher's guide. Portsmouth, NH: Heinmann.

Dickinson, S. (2011). *GLM repeated measures.* Retrieved from http://www.indiana.edu/~iscc/RM_ANOVA.html

Entwistle, N. (2007). Conceptions of learning and the experience of understanding: Thresholds, contextual influences, and knowledge objects. In S. Vosniadou, A. Baltas, X. Vamvakoussi (Eds.), *Reframing the conceptual change approach in learning and instruction.* (pp. 123-143). Amsterdam, The Netherlands: Elsevier.

Fang, Z. (2004). Scientific literacy: A systematic functional linguistics perspective. *Science Education*, *89*(2), 335-347.

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*(5), 491–520.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160.

Fradd, S. H., Lee, O., Sutman, F. X., & Saxton, M. K. (2001). Promoting science literacy with English language learners through instructional materials development: A case study. *Bilingual Research Journal, 25*(4), 417-439.

Gee, J. P. (2005). Language in the science classroom: Academic social languages as the heart of school-based literacy. In R. Yerrick & W.M. Roth (Eds.), *Establishing scientific classroom discourse communities: Multiple voices of teaching and learning research* (pp. 19-37). Mahwah, NJ: Lawrence Erlbaum Associates.

Gunel, M., Hand, B., & McDermott. A. (2009). Writing for different audiences: Effects on high-school students' conceptual understanding of biology. *Learning and Instruction, 19*(4), 354-367.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: literary and discursive power*. London, England: Falmer Press.

Halldén, O. (1999). Conceptual change and contextualization. In W. Schnotz, S. Vosniadou, & M. Carretero (Eds.), New perspectives on conceptual change, (pp. 53-66). Oxford, England: Pergamon.

Hand, B., Gunel, M., & Ulu, C. (2009). Sequencing embedded multimodal representation in a writing to learn approach to the teaching of electricity. *Journal of Research in Science Teaching, 46*(3), 225-247.

Hanson, S. L. (1996). Lost talent: Women in the sciences. Philadelphia, PA: Temple University Press.

Herrell, A. I, & Jordan, M. (2008). *50 strategies for teaching English language learners.* Upper Saddle River, NG: Merrill/Prentice Hall.

Huck, S. W. (2008). *Reading statistics and research.* Boston, MA: Pearson Education.

Hudicourt-Barnes, J. (2003). The use of argumentation in Haitian Creole science class-rooms. *Harvard Educational Review, 73*(1), 73-93.

Janzen, J. (2008). Teaching English language learners in the content areas. *Review of Educational Research, 79*(4), 1010-1038.

Jones, L. R., Mullis, I., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). *The 1990 science report card: NAEP's assessment of fourth, eighth, and twelfth graders.* Washington, DC: Educational Testing Service.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York, NY: Norton.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology, 54*(1), 1-20.

Keys, C. W. (2000). Investigating the thinking processes of eighth grade writers during the composition of a scientific laboratory report. *Journal of Research in Science Teaching, 37*(7), 676-690.

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching, 36*(10), 1065-1084.

Kieffer, M. J., Lesaux, N., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *29*(3), 1168-1201.

Kim, M. C., & Hannafin, M. J. (2011). Scaffolding 6th graders' problem solving in technology-enhanced science classrooms: A qualitative case study. *Instructional Science, 39*(3), 255-282.

Kohlhass, K., Lin, H., & Chu, K. (2010). Disaggregated outcomes of gender, ethnicity, and poverty on fifth grade science performance. *Research in Middle Level Education Online, 33*(7), 1-12.

Langer, J. A., & Applebee, A. N. (1987). *How writing shapes thinking.* Urbana, IL: National Council of Teachers of English.

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*(4), 491-530.

Lee, O., Deaktor, R. A., Hart, J. E., Cuevas, P., & Enders, C. (2005). An instructional intervention’s impact on the science and literacy achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching, 42*(8), 857-887.

Lee, O., & Fradd, S. H. (1996). Interactional patterns of linguistically diverse students and teachers: Insights for promoting science learning. *Linguistics and Education: An International Research Journal, 8*(3), 269-297.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with nonmainstream students. *American Educational Research Journal, 42*(3), 411-438.

Lee, O., Maerten-Rivera, J., Penfield, R. D., LeRoy, K., & Secada, W. G. (2008). Science achievement of English language learners in urban elementary schools: *Results of a First-Year Professional Development Intervention. Journal of Research in Science Teaching, 4*(1), 31-52.

Lee, O., Mahotiere, M., Salinas, A., Penfield, R. D., & Maerten-Rivera, J. (2009). Science writing achievement among English Language Learners: Results of three-year intervention in urban elementary schools. *Bilingual Research Journal, 32*, 153-167.

Lemke, J. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex Publishing Corporation.

Li, M., Ruiz-Primo, M. A., Ayala, C., & Shavelson, R. J., (2000, April). *Study of the reliability and validity of inferring students' understanding from their science journals.* Paper presented at the AERA Annual Meeting. New Orleans, L.A.

Lindholm-Leary, K. J. (2001). *Impact of two-way bilingual elementary programs on students' attitudes toward self and college.* Center for Research on Education, Diversity, and Excellence. Santa Cruz, California.

Liu, O. L., Lee, H., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching, 48*(9), 1079-1107.

Luft, J. (1998, January). *Rubrics: Design and use in science teacher education.* Paper presented at the Annual Meeting of the Association for the Education of Teachers in Science, Minneapolis, MN.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching, 42*(8), 912-946.

Lyukx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teachers College Record, 109*(4), 897-926.

Manz, P. H., Fantuzzo, J. W., & Power, T. J. (2004). Multidimensional assessment of family involvement among urban elementary students. *Journal of School Psychology, 42*(6), 461-475.

Mason, L. (2001). Introducing talk and writing for conceptual change: a classroom study. *Learning and Instruction, 11*, 305-329.

Mason, L., & Boscollo, P. (2000). Writing and conceptual change. What changes? *Instructional Science, 28*, 199-226.

Merino, B. & Scarcella, R. (2005). Invited essay: Teaching science to English learners. *University of California Linguistic Minority Research Institute Newsletter, 14*(4), 1-7.

Muller, P. A., Stage, F. K., & Kinzie, J. (2001). Science achievement growth trajectories: Understanding factors related to gender and racial differences in precollege science achievement. *American Educational Research Journal, 38*(4), 981-1012.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 math report card for the nation and the states.* Washington, DC: Government Printing Office.

National Center for Education Statistics. (2011). *The nation's report card: Science 2009* (NCES 2011–451). Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2012). *The condition of education 2012*. Washington, DC. U.S. Department of Education.

National Council of Teachers of English. (2008). *English language learners: A policy brief produced by the national council of teachers of English*. Urbana, Illinois: The James R. Squire Office for Policy Research.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224-240.

Pallant, J. (2010). *SPSS survival manual: A step by step guide to using SPSS.* Berkshire, England: Open University Press.

Panferov, S. (2010). Increasing ELL parental involvement in our schools: Learning from the parents. *Theory Into Practice, 49*(2), 106-112.

Patterson, E. W. (2001). Structuring the composition process in scientific writing. *International Journal of Science Education, 23*(1), 1-16.

Prain, V., & Hand, B. (1996). Writing for learning in secondary science: Rethinking practices. *Teaching and Teacher Education, 12*(6), 609-626.

Pinker, S. (1994). The language instinct. New York, NY: Morrow.

Raykov, T., & Marcoulides, G. A. (2011). Introduction to psychometric theory. New York. NY: Routledge.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). Measurement and assessment in education. In A. E. Burvidovs (Ed.), *Validity for Teachers* (pp. 117-140). Boston, MA: Pearson Education, Inc.

Riegle-Crumb, C., Moore, C., & Ramos-Wada, A. (2011). Who wants to have a career in science or math? Exploring adolescents’ future aspirations by gender and race/ethnicity. *Science Education, 95*(3), 458-476.

Ritchie, S. M., Tomas, L., & Tones, M. (2011). Writing stories to enhance scientific literacy. *International Journal of Science Education, 33*(5), 685-707.

Rivard, L. P., & Straw, S. B. (2000). The effect of talk and writing on learning science:

An exploratory study. *Science Education, 84*(5), 566-593.

Rosebery, A. S., Warren, B., & Conant, F. R. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *The Journal of the Learning Sciences 2*(1), 61–94.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41-53.

Ruiz-Primo, M. A., Li, M., Ayala, C., &Shavelson, R. J. (1999, March). *Student science journals and the evidence they provide: classroom learning and opportunity to learn.* Paper presented at the NARST Annual Meeting. Boston, MA.

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education, 26*(12), 1477–1506.

Ruiz-Primo, M. A., Li, M., & Shavelson, R. J. (2002). Looking into students' science notebook: What do teachers do with them? *National Center for Research on Evaluation, Standard, and Student Testing (CRESST).* CSE Technical Report 562: Stanford University.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Ruiz-Primo, M. A., Li, M., Tsai, S. P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching, 47*(5), 583-608.

Ryoo, K. (2010). *Learning science, talking science: The impact of technology-enhanced curriculum on students' science learning in linguistically diverse classrooms*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Scarcella, R. (2003). *Academic English: A conceptual framework.* Technical Reports, University of California Linguistic Minority Research Institute, UC Berkeley. http://escholarship.org/uc/item/6pd082d4.

Schroeder, C. M., Scott, T. P., Tolson, H., Huand, T., & Lee, Y. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436-1460.

Stoddard, T., Pinal, A. Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. *Journal of Research in Science Teaching, 39*(8), 664-687.

Sutton, C. (1996). The scientific model as a form of speech. In Welford, G., Osborne, J. and Scott, P. (Eds.), *Research in Science Education in Europe* (pp. 143-152). London, England: Palmer Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* Pearson/Allyn & Bacon.

Teachers of English to Speakers of Other Languages (1997). *ESL standards for pre-K–12 students.* Alexandria, VA: Author.

TELPAS, Texas English Language Proficiency Assessment System: Manual for Raters and Test Administrators Grades K-12 (Spring, 2011). Texas Education Agency.

Texas Education Agency. (2010). Texas education agency 2009-10 state performance report. Retrieved from http://ritter.tea.state.tx.us/perfreport/aeis/2010/state.html

Texas Education Agency. (2011a). *Enrollment in Texas public schools, 2010-11.* (Document No. GE12 601 01). Austin TX: Author.

Texas Education Agency. (2011b). Texas education agency 2010-11 state performance report. Retrieved from http://ritter.tea.state.tx.us/perfreport/aeis/2011/state.html

Texas Education Agency. (2011c). *Texas Essential Knowledge and Skills.* Retrieved from http://www.tea.state.tx.us/

Texas Education Agency. (2011-2012). *LPAC, Language Proficiency Assessment Committee Framework Manual.* Austin TX: Author.

Thompson, B. (2008). *Foundations of behavioral statistics: An insight-based approach.* New York, NY: Guildford Press.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). Tool and symbol in child development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Waxman, H. C. (2002). *Research synthesis on effective teaching practices for English Language Learners.* Philadelphia, PA: Mid-Atlantic Lab for Student Success.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3/4), 139-159.

Yore, L. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education, 25*(6), 689-725.

APPENDIX A

JOURNALS SEARCHED

Table A-1

Journals Searched, Their Impact Factors, and Rational

Journal Title	Impact Factor (2011)	Rational
Reading and Writing	3.850	for theory/research
Learning and Instruction	3.727	for empirical research
Review of Educational Research	3.169	for synthesis
Journal of Research in Science Teaching	2.639	for empirical research
American Educational Research Journal	2.393	for empirical research
Instructional Science	1.828	for empirical research
Science Education	1.775	for empirical research
Journal of Educational Research	1.486	for empirical research
International Journal of Science Education	1.050	for empirical research
Elementary School Journal	0.870	for fifth grade studies
Harvard Educational Review	0.700	for theory
Journal of Adolescent and Adult Literacy	0.670	for junior high studies
Urban Education	0.557	for low-SES studies
Bilingual Research Journal	none	for ELL studies

APPENDIX B

LITERATURE REVIEW MATRICES

Table B-1

Summary of Findings from Studies Examining Non-ELL Writing in Science and Conceptual Understanding (N=9)

Study/ Context	Method/ Sample	Writing Use	Findings
(Rivard & Straw, 2000) French-Canadian province (students speak French as first language)	Quantitative Quasi-experimental Treatment and Control groups	**CONVENTIONAL** Writing an explanation - Written prompt during discussion to "scaffold metagconitive awareness during the explanatory session" (p. 575) – writing an explanation to a problem solving task	- Talk is important for sharing, clarifying, and distributing knowledge. - Analytical writing is a tool for transforming ideas into more coherent and structured knowledge. - Talk combined with writing enhances retention of science learning over time.
(Patterson, 2001)* Southern Wales.	Qualitative Years 2, 3 &6 students from different schools; students selected by the teachers (N = 6)	**CONVENTIONAL** Concept maps before writing descriptions and/or explanations about the science topic: ideas scaffolded.	"The process of writing can enhance pupils' learning in science", and providing scaffolds allows students to "demonstrate far greater of concept understanding in their writing" (p. 15).
(Gunel, Hand, & McDermott, 2009)* Midwestern U.S.A. public schools - upper/middle class.	Quantitative Quasi-experimental, pre-posttest design ANCOVA & MANCOVA Twenty 9th grade students and 98 10th grade students in four different classes	**CONVENTIONAL** Writing explanation of nervous system, circulatory system, and respiratory system to *different audiences*: language scaffolded.	Students in Phase 2 who wrote to a 3rd and 4th grade audience scored significantly higher on the conceptual questions (i.e., essay) of the posttest than the students who wrote to the teacher.

Table B-1 Continued

Study/ Context	Method/ Sample	Writing Use	Findings
(Hand, Gunel, & Ulu, 2009) Istanbul, Turkey. Semi-private boarding high school	Quantitative Quasi-experimental, pre-posttest design. ANCOVA Cohen's d	**CONVENTIONAL** Writing explanation letter to 9th grade students; *embedding text with math and graphs*: ideas scaffolded	Statistical mean differences on test measures supported the pattern of advantage of embedding text plus mathematical representations in writing.
(Mason, 2001) Northern Italy; homogenous, middle class background.	Qualitative 12 fourth grade students	**CREATIVE** Promoting peer collaboration (form of scaffold) with talk and reflective writing before, during, and after science activities – students reflecting on their understanding.	Students advanced conceptually at different levels of scientific understanding.
(Mason & Boscollo, 2000) Northern Italy; homogenous, middleclass background	Mixed-Methods Quantitative pre and post science test measures on content understanding (open-ended questions). ANOVA Qualitative analysis of experiment groups' writing. 36 fourth graders (n= 16 in experimental; n= 20 in control) from two public elementary schools in	**CREATIVE** Teacher modeling: reflection writing scaffolded – reflect, reason, and compare (not "canonical")	Quantitative evidence: Students in the experimental group reached higher levels on all the post science test measures than did control students. Qualitative evidence: Experimental group children's written tests show that writing "helped them to better understand the new topic and to refine their metaconceptual awareness of the changes occurred in their own conceptual structures" (p. 222).
(Ritchie, Tomas, & Tones, 2011) "Well-resourced suburban Australian school"(p. 690).	Quantitative Treatment and control; pre-posttest MANOVA; t-tests. 55 6th grade students; Treatment class, n = 28; Comparison class, n = 27	**CREATIVE** Narrative merging science info. with story line: narrative scaffolded by computer program	Students in treatment group showed a significant improvement on posttest scores compared to control & significant differences between the mean scientific content scores from the "pre-test" and "post-test" narratives.

Table B-1 Continued

Study/ Context	**Method/ Sample**	**Writing Use**	**Findings**
(Keys, Hand, Prain, & Collins, 1999) Southeastern U.S.A. small town; 70% white.	Qualitative Two eighth-grade classes; 19 target students.	**MIX** Structured lesson with final genre project – group discussion and explanation – lab report, persuasive essay, multimedia representation, newspaper article; but also writing reflections during process.	Students' engaged in meta-cognitive thinking = understanding.
(Akkus, Gunel, & Hand, 2007) Context not given.	Quantitative Treatment-Control/pre-posttest. ANOVA & ANCOVA 592 students in grades 7-11 (270 control and 322 treatment).	**MIX** Structured lesson with written reflection	SWH groups scored significantly higher than students in the control group iff the teacher provided quality implementation of the approach.

Note: Studies are organized first, by "Writing Use" and second, by date (earliest to latest).

Table B-2

Summary of Findings From Studies Examining ELL Writing in Science and Conceptual Understanding (N=4)

Study/ Context	Method/ Sample	Writing Use	Findings
(Amaral, Garrison, & Klentschy, 2002)* Southern California (U.S.A.) public schools – low SES	Post-hoc *Descriptive statistics.* ELL students enrolled in school district for the full four years. N = 615 in fourth grade; N = 635 in sixth grade.	**MIX with Science Notebooks** Science notebooks used to "collect, record, analyze, and report data for each of the inquiry units" (p. 224). Purpose of notebooks: to develop cognitive knowledge and English language skills.	Students who participated in the district science program for all four years showed that all of the students, including ELLs had higher achievement scores in science the longer they were exposed to the program.
(Fradd, Lee, Sutman, & Saxton, 2001) U.S.A. public schools – low SES	Pre-posttest *Descriptive Statistics* 4th students participating in two projects aimed at developing science materials for ELLs: The Promise Project and the Science for All Project.	**MIX** Students using writing to report within structured inquiry. "Record what you did so others can learn. Consider different ways to express your information" (p. 491).	Students participating in the program showed statistically significant gains in post unit science test scores.
(Lee, Deaktro, Hart, Cuevas, & Enders, 2005)* U.S.A. public urban schools – low SES	Pre-posttest 3rd and 4th grade students in six participating elementary schools (N = 1,500)	**MIX** Students asked to write expository paragraphs or narrative stories describing the science process under investigation within science inquiry units.	Students demonstrated statistically significant gains and large effect magnitudes on all measures of science achievement at the end of the school year.

Table B-2 Continued

Study/ Context	**Method/ Sample**	**Writing Use**	**Findings**
(Lee, Maerten-Rivera, Penfield, LeRoy, & Secada, 2008) U.S.A. public urban schools – Low SES	Pre-posttest HLM 1,134 third grade students in treatment school; 959 third grade students in comparison school.	**MIX** Students writing expository paragraphs to describe science process, explanation, or conclusion + responses to prompts.	Students in the treatment group displayed a statistically significant increase in science achievement. There was no statistically significant difference in achievement gains between ELL students and students who had exited from ELL or had never been ELL.

Note: Studies are organized first, by "Writing Use" and second, by date (earliest to latest).

Table B-3

Summary of Findings from Studies Examining Non-ELL Writing in Science and Academic Language (N=2)

Study/ Context	**Method/ Sample**	**Writing Use**	**Findings**
Patterson (2001)* Southern Wales.	Qualitative Years 2, 3 &6 students from different schools; students selected by the teachers (N = 6)	**CONVENTIONAL** Concept maps before writing descriptions and/or explanations about the science topic: ideas scaffolded.	The use of context maps increased the quantity of writing as well as the use of connective words and number of explanations vs. descriptions on the writing (i.e., quality).
Ruiz-Primo, Li, Ayala, & Shavelson *N*(2004)* Bay Area, California, U.S.A.	Quantitative T-tests; Descriptive statistics. Rubric developed and used to rate science notebooks on “quality of communication” and “understanding”. Six fifth-grade classrooms. (N = 72 science notebooks randomly selected).	**MIX (Science Notebooks)** Science notebook writing integrated into science inquiry units (FOSS). Entries included mix of conventional and creative genres.	Quality of communication did not improve over time, likely due “to the fact that no teacher feedback was found in any of the students’ notebooks”.

Note: Studies are organized first, by “Writing Use” and second, by date (earliest to latest).

Table B-4

Summary of Findings from Studies Examining ELL Writing in Science and Academic Language (N=3)

Study/ Context	**Method/ Sample**	**Writing Use**	**Findings**
(Amaral, Garrison, & Klentschy, 2002)* Southern California (U.S.A.) public schools – low SES	Post-hoc Descriptive statistics. ELL students enrolled in school district for the full four years. N = 615 in fourth grade; N = 635 in sixth grade.	**MIX (Science Notebooks)** Science notebooks used to "collect, record, analyze, and report data for each of the inquiry units" (p. 224). Purpose of notebooks: to develop cognitive knowledge and English language skills.	Students who participated in the district science program for all four years showed that all of the students, including ELLs had higher achievement scores in writing the longer they were exposed to the program.
(Lee, Deaktor, Hart, Cuevas, & Enders, 2005)* U.S.A. public urban schools – low SES	Pre-Posttest 3rd and 4th grade students in six participating elementary schools (N = 1,500)	**MIX** Students asked to write expository paragraphs or narrative stories describing the science process under investigation within science inquiry units.	Students demonstrated statistically significant gains and large effect magnitudes in writing achievement measures at the end of the school year.
(Lee, Mahotiere, Salinas, Penfield, & Maerten-Rivera, 2009) U.S.A. public urban schools – low SES; diverse	Pre-Posttest HLM & HGLM 683third grade students in year 1; 661 students in year 2; and 676 students in year 3.	**MIX** "Students write expository paragraphs describing the scientific process under investigation, explanations and conclusions of science experiments…and responses to the writing prompts provided as supplementary materials" (p. 156).	Students made significant achievement gains each year for form but not content. Students classified ESOL had lower form and content scores than students who had exited from ESOL programs or had never been in ESOL programs.

N

Note: Studies are organized first, by "Writing Use" and second, by date (earliest to latest).

Table B-5

Summary of findings from studies examining relationship between language and concept in science writing (N=5)

Study/ Context	Methods/ Sample	Writing Use	Findings
(Keys, 1999) Southeastern U.S.A.; urban area.	Qualitative Content analysis of written documents produced in a naturalistic setting. 34 students (33 African American; 1 Latino); grades 7-10; from fives schools (4 low SES; 1 middle class).	**CONVENTIONAL** Students participated in inquiry projects; worked with partner to compose written reports about their observations (no explicit writing instruction given).	Student reports contained only a few meaningful inferences; a few students produced linguistic patterns that expressed expansion and generation of scientific ideas.
(Gunel, Hand, & McDermott, 2009)* Midwestern U.S.A. public schools - upper/middle class.	Quasi-experimental, pre-posttest design with two consecutive phases. Regression analysis. Twenty 9th grade students and 98 10th grade students in four different classes. Pre and posttests included multiple choice or true/false questions and an essay question. Writing assignments collected and graded with instructor-created rubrics. ANCOVA & MANCOVA	**CONVENTIONAL** Writing-to-learn activities with instruction and feedback. Phase 1 – all students (four treatment groups) wrote an explanation. Phase 2 - each treatment group wrote a description for a different audience. Class 1 – to 3rd and 4th grades; Class 2 – parents; Class 3 – peers; Class 4 – teacher.	For both Phase 1 and 2, writing assignment scores were significant predictors for performance for posttests.
(Keys, 2000) Rural middle school in Southeaster U.S.A.	Qualitative 16 eighth graders chosen; all of European American decent.	**MIXTURE** Use of the Science Writing Heuristic (SWH) instructional strategy in context of laboratory experiment. Students wrote report with descriptions, analysis, and reflection.	Students who focused on problem solving generated higher levels of scientific thinking and learned science from writing.

Table B-5 Continued

Study/ Context	Methods/ Sample	Writing Use	Findings
(Ruiz-Primo, Li, Ayala, & Shavelson, 2004)* Bay Area, California, U.S.A.	Quantitative T-tests; Descriptive statistics. Rubric developed and used to rate science notebooks on "quality of communication" and "understanding". Six fifth-grade classrooms. (N = 72 science notebooks randomly selected).	**MIXTURE (Science Notebooks)** The classrooms used as science inquiry curriculum as part of a larger study.	Students' science notebook scores can serve as an achievement indicator since they correlated positively with other science measures. "Science notebooks can assist students' thinking, reasoning, and problem solving if used appropriately" (p. 1501).
(Ruiz-Primo, Li, Tsai, & Schneider, 2010) Eight different schools in five states	Quantitative Correlational; Pearson r. N= 72 notebooks from middle school classrooms	**MIXTURE (Science Notebooks)** Science inquiry curriculum used in the classrooms. Beyond this, teachers were only encouraged (not trained or required) to use science notebooks.	Level of student understanding was consistent with the quality of students' explanations.

Note: Studies are organized first, by "Writing Use" and second, by date (earliest to latest).

APPENDIX C

PROJECT MSSELL GRADE 5 LESSON PLAN SEQUENCE EXCERPT

PHYSICAL SCIENCE	
Week 3: Changing Forms of Energy and Electricity	
TEKS	Science 5.6A-C The student knows that energy occurs in many forms and can be observed in cycles, patterns, and systems. • 5.6A The student is expected to explore the uses of energy, including mechanical, light, thermal, electrical, and sound energy. • 5.6B The student is expected to demonstrate that the flow of electricity in circuits requires a complete path through which an electric current can pass and produce light, heat, and sound. • 5.6C The student is expected to demonstrate that light travels in a straight line until it strikes an object or travels through one medium to another and demonstrate that light can be reflected such as the use of mirrors or other shiny surfaces and refracted such as the appearance of an object. ELPS 1A, 1E, 2C, 3D, 4A, 4J, 5B
Objectives	The student will: • demonstrate that light can be reflected using a reflection design • demonstrate that light can be refracted using an investigation and journal • demonstrate that electricity can flow in a circuit using an investigation and journal • demonstrate how an electromagnet works using an investigation and journal
Target Vocabulary	Verbs: describe, identify, demonstrate Content: simple system, interactions, reflection, refraction, electricity, circuit
EARTH/SPACE SCIENCE	
Week 13: Earth in Space	
TEKS	Science (5.5A, 5.5B, 5.6A 1998), (5.8C 2010) 5.5 (1998) Science concepts. The student knows that a system is a collection of cycles, structures, and processes that interact. • 5.5A (1998) The student is expected to describe some cycles, and processes found in a simple system. • 5.5B (1998) The student is expected to describe some interactions that occur in a simple system. 5.6 (1998) Science concepts. The student knows that some change occurs in cycles. • 5.6A (1998) The student is expected to identify events that describe changes that occur on a regular basis such as daily, weekly, lunar, and seasonal cycles.

	5.8 (2010) Earth and space. The student knows that there are recognizable patterns in the natural world and among the Sun, Earth, and Moon system. • 5.8C (2010) The student is expected to demonstrate that Earth rotates on its axis once approximately every 24 hours causing the day/night cycle and the apparent movement of the Sun across the sky. ELPS 1A, 1E, 2C, 3D, 4A, 4J, 5B
Objectives	The student will: • identify describe the rotation of Earth using an illustration • identify the cause of day and night of Earth using a foldable • describe the revolution of Earth around the Sun using an illustration • describe the seasons using a foldable
Target Vocabulary	Verbs: identify, describe Content: axis, rotation, cycle, revolution (interact, daily cycle, seasonal cycle – indirectly taught)
Week 14: Earth and the Moon	
TEKS	Science 5.5A-B, 5.6A (1998) 5.5 (1998) Science concepts. The student knows that a system is a collection of cycles, structures, and processes that interact. • 5.5A (1998) The student is expected to describe some cycles, structures, and processes found in a simple system. • 5.5B (1998) The student is expected to describe some interactions that occur in a simple system. 5.6 (1998) Science concepts. The student knows that some change occurs in cycles. • 5.6A (1998) The student is expected to identify events that describe changes that occur on a regular basis such as in daily, weekly, lunar, and seasonal cycles. ELPS 1A, 1E, 2C, 3D, 4A, 4J, 5B
Objectives	The student will: • identify describe the revolution of the moon using an illustration • identify four phases of the moon using a foldable • identify and describe phases of the moon using an accordion book • identify the cause of tides and describe tides using journal
Target Vocabulary	Verbs: identify, describe Content: axis, cycle, lunar cycle, revolution, tides (pattern, interact, event – indirectly taught)
LIFE SCIENCE	
Week 25: How Organisms Survive	
TEKS	Science 5.9A, 5.9C, 5.10A 5.9 Organisms and environments. The student knows that there are relationships, systems, and cycles within environments. • 5.9A The student is expected to observe the way organisms live and survive in their ecosystems by interacting

	with the living and non-living elements. • 5.9C The student is expected to predict the effects of changes in ecosystems caused by living organisms, including humans, such as the overpopulation of grazers or building of highways. 5.10 Organisms and environments. The student knows that organism undergo similar life processes and have structures that help them survive within their environments. • 5.10A The student is expected to compare the structures and functions of different species that help them live and survive such as hooves on prairie animals or webbed feet in aquatic animals. ELPS 1A, 1E, 2C, 3D, 4A, 4J, 5B
Objectives	The student will: • identify and explain the role of organisms in their environment using a food web • describe how living organisms modify their environment using an illustration • describe environmental changes and how they affect organisms using a foldable • identify and describe structural adaptations of organisms that help them survive using a foldable • explain how adaptations help organisms survive using an animal mask
Target Vocabulary	Verbs: describe, compare, analyze, predict Content: environmental changes (thrive, become ill, perish), modify, physical environment, adaptations (structural and behavioral), niche, structure, function, camouflage, mimicry, migration, hibernation, predator, prey

APPENDIX D

SCIENCE NOTEBOOK RURBIC

Code	Type of Entry	1		2		3		4	
		LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT
1	Defining	Definition is incomplete, not understandable. *Example: Mixture. When you put…* AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that significantly hinder or prevent understanding. *Example: Mixture. Wen you put too.*	The definition is incorrect (i.e., connections between academic vocabulary and ideas are incorrect).	Definition is complete but does not have technical terms when appropriate; *Example: Mixture. When you put two or more things together.* AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that may hinder or prevent understanding. *Example: Mixture. When you put two or more in one.* .	The definition is partially correct (i.e., connections between academic vocabulary and ideas are partially correct).	Definition is complete AND has some **technical terms** if appropriate; if picture is included. *Example: Mixture. When you put two or more **materials** together.* AND/OR Occasional errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax); but errors do not hinder or prevent understanding. *Example: Mixture. Wen you put tow or more **materials** together.*	The definition is correct (i.e., connections between academic vocabulary and ideas are correct).	Definition is complete AND has all **technical terms** if appropriate. *Mixture. **Combining** two or more **materials** together forms a mixture.* AND/OR Minimal errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that rarely interfere with communication. *Mixture. **Combine** two or more **materials** together form a mixture.*	The definition is correct AND elaborates on the concept with an example (i.e., the example makes reference to previous or current knowledge).

Code	Type of Entry	1		2		3		4	
		LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT
2	Illustrating and Labeling Diagrams	Illustration or diagram is not identifiable. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that significantly hinder or prevent understanding.	The diagram demonstrates no relationship between concepts and/or descriptions (i.e., label is matched incorrectly; written descriptions are incorrect).	- Illustration or diagram can be easily identified - BUT most of the important parts are not labeled; - May or may not have a title - AND may or may not have technical terms if appropriate. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that may hinder or prevent understanding.	The diagram demonstrates some relationship between concepts and/or descriptions (i.e., some labels are matched incorrectly; written descriptions are partially correct).	- Illustration or diagram can be easily identified - AND most of the important parts are labeled; - Includes a title but no technical terms - OR does not include a title but does have technical terms. AND/OR Occasional errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax); but errors do not hinder or prevent understanding.	The diagram demonstrates adequate relationship between concepts and/or descriptions (i.e., labels are matched correctly; written descriptions are correct).	- Illustration or diagram can be easily identified - AND all of the important parts of the representation are labeled, - AND it has a title, - AND it has technical terms if appropriate. AND/OR Minimal errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that rarely interfere with communication.	The diagram demonstrates adequate relationships between concepts and/or descriptions (i.e., labels are matched correctly; written descriptions are correct) AND demonstrates connections to previously learned concepts/ knowledge.

Code	Type of Entry	1		2		3		4	
		LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT
3	Organizing information using two-dimensional figures (i.e., charts, tables, graphs, schematics).	Figure is not clearly identifiable. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that significantly hinder or prevent understanding.	The figure demonstrates incomplete and/or inaccurate data AND any written notes are incorrect.	- Figure is clearly a table, a graph, or a visual representation, - BUT is not labeled properly (*for example, columns and rows are not labeled*), - May or may not have a title - AND may or may not have technical terms if appropriate. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that may hinder or prevent understanding	The figure demonstrates most of the data BUT some of it is inaccurate AND any written notes are partially correct.	- Figure is clearly a table, a graph, or a visual representation - AND is labeled properly, - Includes a title but no technical terms - OR does not include a title but does have technical terms. AND/OR Occasional errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax); but errors do not hinder or prevent understanding.	The figure demonstrates most of the data AND the data demonstrated is accurate AND any written notes are correct.	- Figure is clearly a table, a graph, or a visual representation, - AND is labeled properly, - AND has a title - AND has technical terms if appropriate. AND/OR Minimal errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that rarely interfere with communication.	The figure demonstrates all of the data AND the data demonstrated is accurate AND any written notes are correct AND make connections to previous knowledge and/or demonstrate student reflection/conclusions based on new knowledge.

Code	Type of Entry	1		2		3		4	
		LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT
4	Recording observations and predictions	Entry is incomplete. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that significantly hinder or prevent understanding.	The observations or predictions demonstrate no connections between what is observed and what is written/ predicted.	- Observation or prediction is present, - BUT is not logical (*i.e., observation does not match with topic/ prediction does not include a reason for the prediction*) - AND it may or may not have technical terms if appropriate. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that may hinder or prevent understanding.	The observations or predictions demonstrate partial connections between what is observed and what is written/ predicted (i.e., predictions provide some logical justifications based on previous learning).	- Observation or prediction is present, - AND is somewhat logical (*i.e., observation mostly matches with topic/ prediction includes a plausible reason for the prediction*) - AND it may or may not have technical terms if appropriate AND/OR Occasional errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax); but errors do not hinder or prevent understanding.	The observations or predictions demonstrate connections between what is observed and what is written/ predicted (i.e., predictions include logical justifications based on previous learning).	- Observation or prediction is present, - AND observation is logical (i.e., observation matches topic/ prediction includes a convincing reason for the prediction) - AND it has technical terms if appropriate. AND/OR Minimal errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that rarely interfere with communication.	The observations or predictions demonstrate connections between what is observed and what is written/ predicted (i.e., predictions include several logical justifications based on previous learning).

Code	Type of Entry	1		2		3		4	
		LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT	LANGUAGE	CONCEPT
5	Reflecting	Entry is incomplete. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that significantly hinder or prevent understanding.	The reflection demonstrates no connections between academic vocabulary and concepts learned.	- Reflection is present, - BUT does not make logical reference/ connection to the topic at hand - AND it may or may not include technical terms if appropriate. AND/OR Frequent errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that may hinder or prevent understanding.	The reflection demonstrates connections between academic vocabulary and concepts learned, BUT the connections are not fully and/or logically justified.	- Reflection is present, - AND makes a somewhat logical reference/ connection to the topic at hand - AND it may or may not include technical terms if appropriate. AND/OR Occasional errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax); but errors do not hinder or prevent understanding.	The reflection demonstrates connections between academic vocabulary and concepts learned, AND the connections are fully and/or logically justified.	- Reflection is present, - AND makes a logical reference/ connection to the topic at hand - AND includes technical terms if appropriate. AND/OR Minimal errors in mechanics (i.e., spelling, punctuation) and grammar (i.e., word order, syntax) that rarely interfere with communication.	The reflection demonstrates connections between academic vocabulary and concepts learned, AND the connections are fully and/or logically justified AND the reflection raises new questions about the concept.

APPENDIX E

RUBRIC MANUAL

Unit 1: Week 3 – Physical Science: Changing Forms of Energy and Electricity

Number of activities in science notebook: 6

Technical Words (Taken from target vocabulary from MSSELL lesson plans based on state standards).

Verbs: describe, identify, demonstrate

Content: simple system, interactions, reflection, refraction, electricity, circuit

Concepts

Activity 1. Target Reflection Activity
    Purpose: to demonstrate that light can be reflected using a reflection design.
Activity 2. How Light Moves
    Purpose: to demonstrate that light can be refracted using an investigation.
Activity 3. Batteries, Bulbs, and Wires
    Purpose: to demonstrate that electricity can flow in a circuit using an investigation.
Activity 4. What Can Electricity Flow Through?
    Purpose: To predict and test which objects electricity can flow through using an investigation.
Activity 5. How are Series and Parallel Circuits Different?
    Purpose: To explore and conclude how series and parallel circuits differ using an investigation.
Activity 6. Can You Change the Poles of an Electromagnet?
    Purpose: To demonstrate how an electromagnet works using an investigation.

Unit 2: Weeks 13-14 – Space Science: Earth in Space; Earth and the Moon

Number of activities in science notebook: 7

Technical Words (Taken from target vocabulary from MSSELL lesson plans based on state standards).

Verbs: identify, describe

Content: axis, rotation, cycle, revolution, interact, daily cycle, seasonal cycle, lunar cycle, revolution, tides, pattern, interact, event

Concepts

Activity 1. Model of Day and Night
    Purpose: To create a model of day and night using everyday objects.
Activity 2. Rotation of Earth on Axis/ Rotation of Earth around Sun
    Purpose: To use a foldable to describe the rotation of the earth on its axis and the rotation of the earth around the sun
Activity 3. Earth's Revolution around the Sun
    Purpose: To describe the revolution of Earth around the Sun using an illustration.
Activity 4. Seasons
    Purpose: To describe the seasons using a foldable and/or illustration.
Activity 5. The Revolution of the Moon
    Purpose: To describe the revolution of the moon using an illustration.
Activity 6. Phases of the Moon
    Purpose: To identify the four phases of the moon using a foldable.
Activity 7. Phases of the Moon Cont.
    Purpose: To identify and describe all of the phases of the moon using an accordion book.

Unit 3: Week 25 – Life Science: How Organisms Survive

Number of activities in science notebook: 5

Technical Words (Taken from target vocabulary from MSSELL lesson plans based on state standards).

Verbs: identify, describe

Content: environmental changes (thrive, become ill, perish), modify, physical environment, adaptations (structural and behavioral), niche, structure, function, camouflage, mimicry, migration, hibernation, predator, prey

Concepts

Verbs: classify

Content: matter, physical properties, physical state (solid, liquid, gas), freezing, melting, condensation, evaporation, magnetism

Activity 1: Roles of Organisms in Their Environment

Purpose: Identify and explain the role of organisms in their environment using a food web.

Activity 2: Illustration

Purpose: to describe how living organisms modify their environment using an illustration.

Activity 3: Warm-up (Habitat Change)

Purpose: To explain how habitat changes can affect animals.

Activity 4: Environmental Change

Purpose: to describe environmental changes and how they affect organisms using a foldable.

Activity 5: Structural Adaptations

Purpose: To explain how adaptations help organisms survive using an animal mask.

APPENDIX F

RUBRIC SCORING FORM

Student Code # ____

Scoring Form for Physical Science Unit (Week 3 –Changing Forms of Energy and Electricity)				
Activity	Complete? y/n	Type of Entry 1-4	Language Score 1-4	Concept Score 1-4
Activity 1. Target Reflection Activity				
Activity 2. How Light Moves (Reflection)				
Activity 3. Batteries, Bulbs, and Wires				
Activity 4. What Can Electricity Flow Through?				
Activity 5. How are Series and Parallel Circuits Different?				
Activity 6. Can You Change the Poles of an Electromagnet?				
Mean Score (Sum of scores ÷ # of complete entries)				

Student Code # ____

Scoring Form for Space Science Unit
(Weeks 13-14 –Earth in Space; Earth and the Moon)

Activity	Complete? y/n	Type of Entry 1-4	Language Score 1-4	Concept Score 1-4
Activity 1. Model of Day and Night				
Activity 2. Rotation of Earth on Axis/ Rotation of Earth around Sun				
Activity 3. Earth's Revolution around the Sun				
Activity 4. Seasons				
Activity 5. The Revolution of the Moon				
Activity 6. Phases of the Moon				
Activity 7. Phases of the Moon Cont.				
Mean Score (Sum of scores ÷ # of complete entries)				

Student Code # ____				
Scoring Form for Life Science Unit (Week 25 – How Organisms Survive)				
Activity	Complete? y/n	Type of Entry 1-4	Language Score 1-4	Concept Score 1-4
Activity 1: Roles of Organisms in Their Environment				
Activity 2: Illustration				
Activity 3: Warm-up (Habitat Change)				
Activity 4: Environmental Change				
Activity 5: Structural Adaptations				
Mean Score (Sum of scores ÷ # of complete entries)				
Grand Mean Score for Language for all Three Units (Sum of Language Scores ÷ 3)				
Grand Mean Score for Concept for all Three Units (Sum of Concept Scores ÷ 3)				