UNIVERSITY OF CALIFORNIA

SANTA CRUZ

UNRAVELING THE COMPLEX: CHANGES IN SECONDARY SCIENCE PRESERVICE TEACHERS' ASSESSMENT EXPERTISE

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

EDUCATION

by

Edward G. Lyon

June 2012

The Dissertation of Edward G. Lyon
is approved:

_____
Professor Jerome M. Shaw, Chair

_____
Professor Doris B. Ash

_____
Professor Trish Stoddart

_____
Professor Jonathan F. Osborne

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

UNRAVELING THE COMPLEX: CHANGES IN SECONDARY SCIENCE PRESERVICE TEACHERS' ASSESSMENT EXPERTISE

by

EDWARD G. LYON

ABSTRACT

Becoming prepared to assess science learning is a daunting endeavor. Expert science teachers assess a wide range of scientific knowledge and practices, modify instruction based on assessment information, and consider the needs of their diverse learners. These skills are informed by but not necessarily consistent with what one believes and knows about assessment. In this study, I explored the ways in which the assessment expertise of 11 secondary science preservice teachers changed during their 12-month teacher education program. Conceptually, I framed assessment expertise along three dimensions: (a) constructing coherently linked assessments (Construction), (b) using assessment to support students' science learning (Use), and (c) equitably assessing English learners (Equity). Informed by the Construction-Use-Equity (CUE) framework, I designed a set of activities to support teacher learning about assessing science in linguistically diverse classrooms and delivered these activities in three of the participants' courses. Given this context, I report on both conceptual development and empirical findings from the study. First, via a scoring rubric, I articulate how the multidimensional CUE framework translates into observable levels of assessment expertise. Next, I report on the mixed methods analysis of surveys, interviews, and teacher products. Broadly, descriptive and non-

parametric statistical analyses indicate that the teachers' assessment expertise changed differently among the three conceptual dimensions. Qualitative analyses describe the nature of these changes. Most notably, the teachers (a) considered the alignment between assessment task and learning objective, (b) expanded their repertoire of assessment tasks and general assessment strategies, (c) moved toward a formative view of assessment, and (d) became increasingly aware about the role of language in assessment. Finally, in greater detail, I explore the latter pattern by incorporating qualitative analysis of three teachers' classroom practices. These resulting case studies illustrate tensions about whether to assess language in addition to science content and whether to scaffold language use in assessment rather than reduce the use of language. By furthering a conceptualization of assessment expertise and by identifying and describing changing expertise, the study aims to inform how teachers are prepared to assess science.

ACKNOWLEDGEMENTS

First, I would like to acknowledge that a University of California, Santa Cruz Summer Dissertation Fellowship and a University of California All Campus Consortium for Research in Diversity (UC/ACCORD) Dissertation Fellowship funded the writing of this dissertation. Furthermore, I have been funded and greatly informed by participating as a graduate student researcher in the NSF funded Effective Science Teaching for English Language Learners (ESTELL) Project (grant #0822402), led by Dr. Trish Stoddart.

There are many individuals who have each have played an important role in making this dissertation possible. First, I thank the 2010/2011 secondary science cohort for participating in this study. Of course, *you* are those ones that remind us of why we do the work that we do. I thank Preetha Menon and Saul Maldonado for scoring teacher products and Eric Romero, Ei Ei Zin, Carolyn Augustin, Michelle Tang, and Neiman Moore for transcribing interviews. I thank Ana England, Geoff Smith, and Dr. Katherine Nielson for providing me access to your students and your classroom. I am indebted to Dr. Doris Ash, Dr. Trish Stoddart, and Dr. Jonathan Osborne for providing me with so much feedback and support on my dissertation. Of course, I thank my advisor, Dr. Jerome Shaw, for the countless emails, conversations, and experiences that prepared me for the world of academia. Finally, thank you to my wife, Adelyn, for being so patient, supportive, and loving throughout the entire process.

INTRODUCTION

THE COMPLEXITY OF SCIENCE CLASSROOM ASSESSMENT

> *Well it's a lot more complex and tricky than it seemed back then. I mean back then it seemed like straight forward; you know you have formative assessment, summative assessment, and give feedback in the rubrics. But it's so complex. I mean everything you do can be an assessment really. And so it's just being able to be organized enough so that you know what assessment you want to pay attention to like what you're doing each day. And organized enough so that you can have all the rubrics ready for the students when they start. Yea, so I think coming to realize like how broad of a definition assessment is and…not just kind of make it up as you go.*
>
> -Lauren, Interview 3

In the quotation above, Lauren, a preservice science teacher and one of my research participants, is reflecting on how her views and knowledge about assessment have changed since the beginning of her teacher education program. Lauren's comments reflect well why I am investigating the topic of science classroom assessment and what I hope to contribute by reporting my findings. This introduction intends to clarify these two points and to outline the structure of my dissertation.

If you were to eavesdrop on a university course related to teaching or schooling and wait long enough, you are likely to hear some debate around assessment. Either or both the instructor or student may be questioning to what extent assessment does a disservice to kids, whether assessment really measures anything useful, or whether assessment steers curriculum in the wrong direction. In classes aimed at helping teachers put educational theory into practice, you might hear some of the terms espoused by Lauren, such as formative assessment and summative assessment. However, those conversations often oversimplify what it means to assess and focus more on issues, rather than on solutions. Although it is important to know

about, for example, the differences between a summative and formative use of assessment, know *how* to assess science in a particular instructional and student context is a critical component of science teaching. Furthermore, this application of assessment knowledge into practice takes considerably more effort.

My experiences as a former high school teacher reflect, in many ways, the struggles preservice science teachers might face. In my teacher education program, I encountered many perfunctory discussions around assessment – never given an opportunity to interrogate my own beliefs or opportunities to reflect critically on how other teachers assess science. I often equated assessment with the task itself, and, as a novice teacher, I would expend an inordinate amount of energy designing assessments that often resembled standardized science tests.

Due to many factors, including my teacher education program's focus on social justice and collaborative efforts with colleagues, I eventually found a space to venture into alternative assessment forms. I explored the use of concept mapping, group oral exams, project-based assessments, and performance assessments. My impetus for trying out new ideas is that I believed that my culturally and linguistically diverse students might not all demonstrate scientific understanding in the same way. Furthermore, I wanted to engage all of my students in complex scientific thinking, even during the assessment.

While these experiences led me to venture into alternative assessment forms, it was not until I enrolled in a doctoral program that I began to see how assessment was so much more than the task. I came to understand assessment as a social activity

marked by intentionality and decisions, instead of just an instrument of evaluation. I came to understand the complexity that Lauren has begun to recognize.

To go beyond, as Lauren puts, the "straightforward," teachers need exposure to various assessment principles and tools. Moreover, teachers need opportunities to consider how to apply such principles and tools in light of *what* science is being taught, *how* that science is being taught, and *who* is being taught. For instance, Lauren, who is obtaining her teaching credential in the state of California, is likely teach some of the 1.5 million students who are non-native English speakers, many of whom identified as English learners, or ELs. In this student context, she faces considerable challenges while assessing science. Perhaps Lauren strongly believes in "science writing" and views lab reports, research papers, or other productive forms of literacy as a way to access the nature of science. Yet, Lauren may still hold onto a belief that her ELs' limited English proficiency would preclude them from being able to write in English well enough to communicate understanding. Thus, even though she might give her students opportunities during instruction to write science, she might be hesitant to assess her students' science writing in fear that the task would be unfair. However, it could be argued that access to such literacy tasks *is* particularly important for ELs. This is just one example of assessment's complexity.

Complexity is not limited to the practice of assessing. Studying science classroom assessment is also a daunting task. A researcher might observe a teacher asking students to dissect flower parts. Is this teacher assessing? How would I know? I would need to go beyond an a priori observational checklist. I would need to

understand the teacher's intentions – why is she having the students dissect flowers? Only then could I infer whether the intention was solely to help students learn about flowers *or* whether another intention was to *find out* what the students know about flowers to support student learning. In a similar vein, if the teacher does find out information about what her students know – by asking questions, interpreting some written responses, or observing them dissect the flower – what is done with such information? Nothing? Is it transformed into a grade? Used to guide possible discussion questions? To study assessment at the classroom level means combining different methods of study and make sense of how teachers think.

To summarize, beyond the critical need to study science classroom assessment, I am intrigued by the complexity of assessing science and studying how science teachers assess. Through conceptual development and an empirical study, my goal is to begin to unravel this complexity. Thus, I had to think about science classroom assessment in two reciprocating ways. First, I had to translate the phenomenon into instruction to help preservice science teachers understand science classroom assessment and do their own personal unraveling. Second, I had to choose appropriate methods in which to collect and analyze teachers' beliefs, knowledge, and facility with science classroom assessment, and to analyze change during their teacher education program.

Within a particular context – a cohort of secondary science preservice teachers exposed to instruction focused on assessing science linguistically diverse classrooms – this study sought to answer the following two research questions:

1. In what ways does the teachers' assessment expertise change during the teacher education program?

2. How do the teachers assess science as part of the Performance Assessment for California Teachers (a culminating teaching event).

To report on both conceptual development and empirical findings, this dissertation is organized into three stand-alone papers, which I refer to as chapters. In Chapter 1, "Unpacking the Complexity in Science Classroom Assessment: Development and Application of the Construction-Use-Equity (CUE) Assessment Expertise Framework," I lay out the Construction-Use-Equity (CUE – pronounced *Q*) Assessment Expertise Framework to frame my conceptualization of assessment expertise. My goal in this chapter is to explicate how this framework productively translates into observable levels of assessment expertise. I provide written exemplars from teacher responses to survey questions and program assignments to instantiate those levels. Furthermore, I argue that a multidimensional approach to analyzing assessment expertise provides a more complete and nuanced picture.

In Chapter 2, "Rough Waters and Smooth Sailing: Charting the Changes of Secondary Science Preservice Teachers' Assessment Expertise," I report on the mixed methods analysis of multiple data sources, including surveys, interviews, and various teacher products. The purpose of this empirical study is to present identified patterns of change through non-parametric tests, descriptive statistics, and qualitative analyses.

Finally, Chapter 3, "What about Language? Case Studies of Secondary Science Preservice Teachers' Evolving Expertise in Equitable Science Assessment," expands on one pattern that emerged from the qualitative analysis: teachers' awareness of the language of assessment. I use this frame of reference to report on three teachers – Dean, Glenda, and Lauren. Beyond a more descriptive account of change, by adding a qualitative analysis of teacher practice, the case studies illustrate successes, struggles, and tensions while enacting equitable assessment understandings in practice.

Collectively, I moved through deepening layers of conceptualization and analysis to unravel what it means to become an expertise in science classroom assessment and unravel changes that occurred as the teachers transitioned toward increasing expertise. This study will set the stage for further explorations that look deeply into issues of secondary science preservice teachers' assessment preparation, particularly in linguistically diverse classrooms. The ultimate hope is that science teachers, such as Lauren, learn how to use assessment for particular functions, select the content and form of assessment carefully, consider the various language demands, and provide supports so that all students can successfully navigate the demands. However, for that to happen researchers need to unravel the complexities of science classroom assessment – and I anticipate that this dissertation will become part of that conversation.

CHAPTER 1

UNPACKING THE COMPLEXITY IN SCIENCE CLASSROOM ASSESSMENT: DEVELOPMENT AND APPLICATION OF THE CONSTRUCTION-USE-EQUITY (CUE) ASSESSMENT EXPERTISE FRAMEWORK

Abstract

Although research in science education has led to new assessment forms and functions, the reality is that little work has been done to unpack the complexities of what it means for a teacher to develop expertise in science classroom assessment. The purpose of this paper is to explicate levels of assessment expertise, both in theory and in practice, as well as to demonstrate how a multidimensional conceptual approach provides a more nuanced and complete picture of science teachers' assessment expertise. First, I describe the conceptual grounding for assessment expertise through the Construction-Use-Equity (CUE) Assessment Expertise Framework. The framework consists of three dimensions – (a) constructing theoretically cohesive assessments that elicit student thinking (Construction), (b) using assessment to support students' science learning (Use), and (c) assessing in ways that are equitable for English learners (Equity). Next, I describe how the framework translates into a rubric used to score preservice teacher responses to various assessment-related prompts. Finally, I provide exemplars from these responses to instantiate the theoretical expertise levels defined in the rubric. The contribution of this paper lies in its further conceptual development of assessment expertise that researchers can use in a host of studies to investigate science classroom assessment.

Introduction

Two decades ago, Wolf, Bixby, Glen III, and Gardner (1991) distinguished between *testing*, where individuals' immutable intelligence is measured with "technically elegant" instruments, and *assessing*, where intelligence is acknowledged as socially constructed and where the assessor seeks to find out how individuals think. In making such as distinction, Wolf et al. called for a transition from a culture of testing to a culture of assessing (p. 33). Since then, scholars have illuminated what an "assessment culture" might look like and how changing views about measurement, learning, and curriculum inform such a paradigm shift (Broadfoot, 1996; Gipps, 1994, 1999; Shepard, 2000).

An important aspect of an assessment culture is that while assessment occurring within the context of classrooms, thus *classroom* assessment, might open up new opportunities for student learning, at the same time, it presents new challenges to successfully enact. As Shepard (2000) argues:

> [C]lassroom assessment must change in two fundamentally important ways. First, its form and content must be changed to better represent important thinking and problem solving skills in each of the disciplines. Second, the way that assessment is used in classrooms and how it is regarded by teachers and students must change. (p. 5)

Science education has begun to address both changes. First, instead of just assessing factual knowledge, seminal science education documents and large-scale science assessments emphasize assessing deep conceptual understanding and scientific practices (Millar & Osborne, 1998; National Center for Education Statistics [NCES], 2009; National Research Council [NCR], 1996, 2012; Organisation for

Economic Co-operation and Development [OECD], 2009). For example, researchers have developed science performance assessments that elicit competencies associated with inquiry-based science in more authentic contexts. Regarding Shepard's (2000) second charge, like other seminal documents, the United States' *National Science Education Standards* (NCR, 1996) calls for multiple assessment purposes, most notably using assessment to support learning. Science education researchers have studied the effect of assessment used to support science learning (i.e., formative assessment) on student outcomes and have described how science teachers have used and thought about formative assessment in the classroom (Bell & Cowie, 2001; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Shavelson et al., 2008). Thus, tenets set in science education documents as well as science education research have embraced the transition into an assessment culture.

Yet, one line of research relatively absent in the science assessment literature relates to teacher preparation. To translate a vision for an assessment culture into reality, science teachers must be prepared so that they can incorporate those assessment strategies emphasized in science education into their classroom practice. As stated by Gipps (1999), a key for future direction in assessment is the "*development* of teachers' classroom assessment skills" and "*development* of new assessment strategies for use by teachers" (p. 387, emphases added). Research in science assessment should follow suit by examining preservice science teachers' assessment beliefs, knowledge, and facility.

While some empirical research does address science teachers' assessment preparation (cf., Siegel & Wissehr, 2011), more work is needed to articulate (a) the theoretical basis on which we study science assessment from a teacher-centered perspective, particularly in light of assessing diverse students, and to articulate the (b) observable instantiations of theory. Given the complexities of science classroom assessment, I argue for a multidimensional approach to study teachers' assessment *expertise* – a construct that I will continue to explicate throughout this paper. I will argue for this approach by first laying out how theories of teaching expertise and pedagogical content knowledge contribute to the development of a conceptual framework, the Construction-Use-Equity (CUE – pronounced *Q*) Assessment Expertise Framework. The CUE framework outlines the complex set of understandings and facilities needed to develop assessment expertise via three conceptual dimensions. Second, I will describe how the CUE framework informed the analysis of written preservice science teacher products using a scoring rubric.

By reporting and commenting on exemplar teacher responses analyzed during an exploratory study of preservice teachers, I seek to illuminate how researchers can study science teachers' assessment expertise and describe how this expertise might look in practice. I conclude by discussing potential next steps in studying science classroom assessment – to unpack further its complexity – and point out implications for science teacher education and science teaching.

Theorizing about Science Classroom Assessment

Teaching Expertise

Becoming an expert in teaching involves more than just acquiring more knowledge about disciplinary content and general pedagogical techniques. Instead, developing expertise entails a shift in teacher thinking, from limited, context-free, and *in*flexible to elaborated, situated, and flexible (Bransford, Brown, & Cocking, 2000; Dreyfus & Dreyfus, 1986). Like all learners, teachers build upon their prior knowledge and experiences, typically resulting in a deeper understanding of related concepts and capacity to apply concepts to the nuances of particular teaching contexts. Such deeper understanding can be captured as a shift from "knowing that" to "know how" (Dreyfus & Dreyfus, 1986; Kuhn, 1970). While knowing "that" is based upon adhering to rules and readily accepted theoretical orientations, knowing "how" is a flexible application of principles in practice to adapt to constantly changing situations.

One characteristic of expertise in any field is a vast domain-specific knowledge base (Glaser & Chi, 1988). However, gaining disciplinary specific expertise is not enough. Although teachers have presumably spent years acquiring content knowledge, Shulman (1986) proposed a central question in studying teacher expertise: "How does the successful college student transform his or her expertise in the subject matter into a form that high school students can comprehend?" (p. 8). To shift toward teaching expertise, Shulman argues that teachers must develop knowledge used to teach others *discipline*-specific content – referred to as pedagogical content knowledge, or PCK. Although acquiring PCK alone does not confer teaching expertise, it is a useful starting point when conceptualizing

assessment expertise. Thus, science teachers' knowledge of assessment is not relegated to general strategies, but rather situated within the disciplinary content being taught.

Assessment Expertise

Magnusson, Krajcik, & Borko (1999) asserted that PCK for science teaching includes knowing *what* to assess and *how* to assess. Building from Magnusson et al., Abell and Siegel (2011) organized assessment PCK into (a) knowledge of *assessment purposes*, (b) knowledge of *what to assess*, (c) knowledge of *assessment strategies*, and (d) knowledge of *interpretation and action taking*. In addition to assessment knowledge, Abell and Siegel argue that at the core are teachers' views of learning and assessment values, described as the "overarching ideas and beliefs that guide assessment decisions in the science classroom" (p. 212). Other researchers have argued for views of learning and assessment as critical to preparing assessment-capable teachers (Black & Wiliam, 1998a; Hill, Cowie, Gilmore, & Smith, 2010) and studying them (Lyon, 2011).

Collectively, both knowledge of assessment principles and views about learning and assessment – in a discipline-specific context – form a foundation in which to generate frameworks that lay out what teachers should be able to do with assessment. Some researchers refer to such frameworks as assessment literacy (Lukin, Bandalos, Eckhout, & Mickelson, 2004; Siegel & Wissehr, 2011; Webb, 2002). Gearhart et al. (2006) studied assessment-related teacher professional development that emphasized the "*relationship* between understandings of assessment concepts

and facility with assessment" (p. 259, emphasis added). They referred to this relationship as assessment expertise, which is consistent with the view that teaching expertise involves knowing *how* to do something, not just acquiring the pedagogical content knowledge needed to teach.

To investigate how preservice science teachers draw on their assessment knowledge and beliefs while assessing student learning, I employed assessment expertise as the focal construct, conceptualized, like Gearhart et al. (2006), as a relationship between assessment understanding (including knowledge and beliefs) and assessment facility (how teachers use such understandings in teaching scenarios and day-to-day planning). One goal of this investigation was to not only lay out exactly what teachers should know and do in the context of science teaching, but also explicate different *levels* of assessment expertise to study *changes* in teachers' assessment expertise. My central question therefore was *in what ways does the assessment expertise of secondary science preservice teachers change?* In this paper, data from the investigation are used to clearly elucidate the types of assessment facilities that the teachers engaged in and how such facilities fall on various expertise levels. Toward this end, I examined assessment expertise through a multidimensional framework, which I describe next. While exploring assessment expertise through multiple dimensions, I came to paint a more complete and detailed picture of science teachers' facility with science assessment. I argue that the framework and developed scoring instrument are helpful in studying science classroom assessment from a teacher-centered perspective.

The CUE Assessment Expertise Framework

Conceptualization of assessment expertise is grounded in sociocultural theory. A sociocultural perspective views learning as occurring through the social interaction between teacher and students and between students (Tharp and Gallimore, 1988; Vygotsky, 1978). In essential ways, assessment plays a role in the learning process, and therefore, is also part of these social interactions. The content being assessed represents the type of knowledge valued in education (Gipps, 1999) and provides a means by which students and teachers communicate with each other about progress toward learning goals. Through a sociocultural perspective, such progress aligns with Vygotsky's (1986) notion of the zone of proximal development (ZPD), or the region between what students can do on their own and what they can do with a more capable peer (see also Wells, 1999). As Haertel, Moss, Pullin, & Gee (2008) summarize, "[A]ssessment and the routines that surround it do far more than provide information – they shape people's understanding about what is important to learn, what learning is, and who learners are" (p. 9).

Grounded in sociocultural theory, the CUE Assessment Expertise Framework is divided into the following dimensions: (a) constructing theoretically cohesive assessments that elicit student thinking, (b) using assessment to support students' science learning, and (c) assessing in ways that are equitable for English learners. The framework suggests a spectrum of assessment expertise aimed at providing a roadmap for developing research instruments, and teacher preparation and

professional development curriculum. The framework also informs the analysis of teacher discourse, products, and practices.

Construction: constructing theoretically cohesive assessments that elicit student thinking. The core aspect of this dimension is that assessment is a *process of inferring about what students know and can do* (NCR, 2001; Popham, 2006). A common design perspective about assessment is that its function is to measure well-defined constructs, given that the assessor does not have direct access to such constructs. In order to draw valid inferences about students' demonstration of the desired construct, the assessment developer would focus on the technical quality of the task – does it measure what it is suppose to measure (validity) and can it be scored consistently (reliability)?

Sociocultural theory helps us recognize that assessment is more than just a measuring tool. Since assessment often drives classroom instruction, it (intended or not) conveys to students what is important to learn and what *display* of knowledge is valued. The inextricable link between learning and assessment means that assessment should draw on how students learn and represent knowledge. The National Research Council report, *Knowing What Students Know: The Science and Design of Education Assessment* (NCR, 2001), takes a psychological approach by asserting that while the *purpose* of assessment can vary from measuring achievement to evaluating programs to assisting learning, any assessment involves three core features, which should be designed as a coordinated whole. The report represents the link among these core features as an assessment triangle (see Figure 1). In the model, *cognition* refers to

theories of how students learn and represent knowledge, *observation* refers to tasks that allow one to observe student learning, and *interpretation* refers to the reasoning one makes from the student's evidence of learning.

The alignment of cognition to observation resonates with Wiggins and McTighe's (2005) understanding by design model in which the teacher must ask, "What will we accept as evidence that the students understand the objective?" In order to assess student learning in science classrooms, the task should also align with appropriate learning objectives. These objectives may relate to disciplinary content, process skills, language development, or attitudes. For instance, the Programme for International Student Assessment (PISA) draws on a framework for scientific literacy, which informed the developed of assessment items used for an international assessment (OECD, 2009). At the classroom level, Magnusson, Krajcik, & Borko (1999) have argued that views (beliefs) of scientific literacy should inform science teachers about what to assess. Science education reform documents emphasize scientific understanding of core ideas, as opposed to rote memorization, as well as engaging in scientific practices – the types of activities that scientists do. Thus, an aspect of expertise in science assessment is not only aligning assessment tasks to what science is being learning and how science is being learned, but also considering what science is being *assessed*.

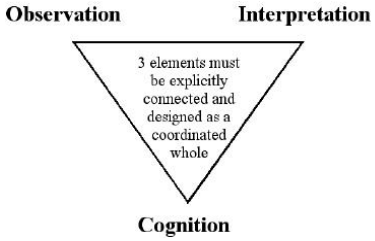

Figure 1. Representation of the assessment triangle model (Pellegrino, Chudowsky, and Glaser, 2001), used to guide the Construction dimension of the CUE framework.

Use: using assessment information to support students' science learning. Instead of focusing on the process of finding out what students know and can do, the Use dimension focuses on what teachers *do* with information gathered by assessing. As described in the introduction, science education emphasizes using assessment to *support* learning (i.e., formative assessment) defined by Black & Wiliam (1998a) as "all those activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities" (p. 140). Black and Wiliam's description emphasizes the function and not the form of the assessment – to be formative, "feedback needs to contain an implicit or explicit recipe for future action" (Wiliam & Leahy, 2007, p. 31). As a way to make sense of the varied functions, even within formative assessment, Ruiz-Primo and Furtak (2007) proposed a formative assessment spectrum from formal (often involving pre-designed curriculum-embedded tasks) to informal (the ongoing interaction between student-teacher and student-student). While understanding and engaging in both sides of the spectrum are valuable aspects of assessment expertise, the Use dimension focuses on the formal side of the spectrum. Specifically,

assessment as a feedback cycle or a cycle of inquiry models how a teacher would cyclically consider: "Where do I want students to get to?" (i.e., the learning objective), "What do students know?" (by interpreting information gathered), and finally "What can be done, given this information, to get them there?" (Bell and Cowie, 2001; McMillan, 2007; Ruiz-Primo & Furtak, 2007).

To demonstrate assessment expertise, as conceptualized in the Use dimension, teachers would first understand *that* assessment should be used in ways outlined throughout this sub-section and second understand *how* to use assessment formatively within their specific discipline. For instance, what conceptions and alternative conceptions should one look for while interpreting information gathered by assessing? What type of feedback will be useful to students? How can instruction be modified to move students toward the learning objective? How can students be involved in assessing themselves and their peers? Ultimately, teachers can engage in a full cycle of inquiry –considering the learning objectives, what information they are getting about students, and what would they do about it.

Equity: assessing in ways that are equitable for English learners. Ultimately, equitable science teaching, as conceptualized in this paper, is about giving all students appropriate *opportunities* to learn science and to demonstrate what they have learned. Although important for all students, a focus on language (text, forms of communication, and ways of participating) is at the heart of equitably assessing English learners (ELs). Previous research demonstrates that math and science assessments items are often biased against ELs due to complex sentence structure,

unfamiliar vocabulary, and cultural references (Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Solano-Flores & Nelson-Barber, 2001). Consistent with these findings, a widely held position in psychological and educational research is that in any content-area assessment, such as science, a student's proficiency in the language of the assessment is also being evaluated (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Thus, a first step to demonstrating assessment expertise is to recognize the ways in which assessment biases ELs. Yet, Gipps and Murphy (1994) argue that "focusing on 'bias' in the test itself has distracted attention from wider equity issues" (p. 27). Due to tracking or perceived ability to engage in complex scientific thinking, ELs are often relegated to less rigorous science courses (Pease-Alvarez and Hakuta, 1992; Valdes, 2001). Such denial of access to rigorous science has been linked to disparities in science achievement (Oakes, 1990).

Guided by a focus on language, equitable science assessment means to provide opportunities for all students, regardless of English language status or proficiency, to demonstrate what they know and can do in science (*fairness*) and, through assessment, provide students with *access* to a challenging and meaningful science curriculum. Teachers can more fairly assess by modifying the language of assessment (Abedi, Hofstetter, & Lord, 2004; Abedi & Lord, 2001; Shaftel et al., 2006) or by scaffolding language through such strategies as contextualized materials, visuals, or sentences frames (Siegel, 2007). Finally, teachers can more fairly assess by

incorporating students' home language, culture, and ways of knowing in the assessment (Solano-Flores & Nelson-Barber, 2001).

Yet, teachers need to be cautious about denying students, particularly ELs, with access to a challenging science curriculum while modifying the language of assessment. Through assessment, teachers can expect students, regardless of language proficiency, to meet appropriate science content and language standards. This means incorporating forms of complex thinking and language use into the assessment. This also means integrating science learning with English language development through assessment (Stoddart, Pinal, Latzke, & Canaday, 2002; Stoddart, Solis, Tolbert, & Bravo, 2010). Integrating science and language in the assessment can address equity by supporting ELs who do not have the academic language needed to access discourse in science and even provide them feedback on language use. Language allows ELs to participate in classroom activity; thereby accessing the rigorous subject matter valued by the community (La Celle-Peterson & Rivera, 1994).

Summary. The CUE framework captures the complexity of science classroom assessment. My argument throughout this paper is that by studying science classroom assessment through multiple dimensions, science education researchers can provide a more nuanced understanding of how science teachers demonstrate assessment expertise. Therefore, I will next turn from theorizing about assessment expertise (what aspects of assessment should we capture while studying science classroom assessment?) to describing how the theory aligned with and informed

research to study how 11 secondary science preservice teachers' (referred hereon as "teachers") assessment expertise changed during their teacher education program.

Linking Theory to Analysis: The CUE Assessment Expertise Rubric and Exemplars

My goal throughout this section is to articulate how a multidimensional approach to science assessment, when linked to appropriate methodology, can be useful for analyzing teachers' written responses to open-ended prompts that solicit their assessment expertise. A rubric used to score the responses best exemplifies this linkage between theory and analysis.

I developed the CUE Assessment Expertise Rubric (see Appendix) for the following reasons: (a) to articulate how the CUE Assessment Expertise Framework translates into observable phenomena, (b) to provide a clear and concise means to elicit various levels of teachers' assessment expertise, and (c) to measure changes in or compare teachers' assessment expertise. The use of rubrics is a common practice in teacher and classroom research (e.g., Stoddart, Pinal, Latzay, & Canaday, 2002) and generally includes a set of common criteria that exist across a continuum of proficiency levels (Luft, 1999). The CUE Assessment Expertise Rubric consists of the three CUE framework dimensions – *Construction, Use,* and *Equity*. I based the various levels of expertise on models of expertise development (Dreyfus & Dreyfus, 1986), in which each level represents an increase in specificity and organization from limited, context-free, and inflexible to expanded, situated, and flexible. For example, in the Construction dimension, an increase in levels corresponds with greater

coherence among the assessment task, learning objective, and evaluative criteria. I developed the rubric iteratively by initially drawing on the CUE framework to develop hypothetical levels of expertise, then comparing the hypothetical levels to patterns found in the coded teacher products, leading to rubric refinement. This approach was appropriate since the rubric was in its initial stages of development (not a prevalidated scoring system).

A major goal was to make the rubric sensitive enough to detect changes in assessment expertise while analyzing teacher products. Although the initial iteration of the rubric consisted of three dimensions – one for each CUE dimension – during scoring it became apparent that each dimension included too many criteria and precluded the rubric from being scored reliably by multiple scorers. For instance, the initial rubric included criteria related to the assessment task itself (was the task eliciting factual knowledge versus scientific thinking?) and criteria related to the alignment of the learning objective, assessment task, and evaluative criteria. Teachers often considered assessments that elicited scientific thinking (representing level three on the rubric), but did not consider the alignment between learning objective and assessment task (representing level one on the rubric). Therefore, instead of just three dimensions, I divided each dimension into two sub-dimensions. The subdivision provided a more nuanced interpretation of assessment expertise.

Table 1

Data Sources

Data source	Description of the prompt	Function

Assessment plan (July 2010; December 2010; May 2011)	(1) Choose one of the following science topics – Mendelian genetics, acids and bases, light and optics, or earthquakes, (2) describe in as much detail as possible how you would assess student learning during this unit, and (3) explain why you would assess this way.	Part of my research study to solicit assessment expertise (in all three dimensions) in a uniform way throughout the teacher education program
Learning Theories final project (August 2010)	Describe three activities to teach a particular science standard and two ways to assess that standard. Include a description of background information needed, importance of the content, and connection with theories of learning	Final project for the *Learning Theories* teacher education program course where teachers are introduced to theories of learning and issues in teaching to diverse students
Performance Assessment for California Teachers (PACT) commentary (May 2011)	Answer a series of prompts related to how you planned the focal learning segment (set of science lessons), including its central focus, theoretical framing, language demands, and collection of assessments	Part of a teacher assessment that must be completed by teachers to complete in some California universities to obtain a teaching credential

To demonstrate what assessment expertise looks like across the various dimensions and how teacher responses were interpreted in light of expertise levels, I selected exemplars for each sub-dimension, moving from lower levels to higher levels of expertise. The exemplars come from the following three data sources: (a) the "assessment plan" open response prompt, (b) the *Learning Theories* final project, and (c) the planning commentary for the Performance Assessment for California Teachers (PACT). Table 1 outlines the description and function of each data source. The exemplars sometimes represent the entire response and sometimes only represent an

excerpt from the response. I conclude the section by analyzing an excerpt from a teacher's *Learning Theories* course final project on all sub-dimensions.

Construction Dimension

The Construction dimension consists of two sub-dimensions – *assessment task* and *alignment.* The *assessment task* sub-dimension represents the extent to which the teacher considers assessment tasks that elicit scientific thinking, consistent with the *National Science Education Standard's* (NCR, 1996) emphasis on assessing rich, well-structured knowledge as opposed to discrete, factual pieces of knowledge. As levels increase, teachers shift from using generic assessment tasks to assessment tasks that elicit factual knowledge to assessment tasks that elicit scientific thinking and ultimately activities that better resemble scientific practices and discourse (Tamir, 1998). Scientific thinking might focus on conceptual understanding, scientific process, or practices, as long as the assessment is open-ended and not looking for a dichotomist correct/incorrect response.

The Construction dimension is not limited to the assessment task. Assessment is a process, and according to the assessment triangle model (NCR, 2001), all three components – cognition, observation, and interpretation – should be aligned and coordinated as a whole. Thus, the *alignment* sub-dimension represents the extent to which the teacher considers all three components of the assessment triangle, which equips them to make better decisions about what students know and ultimately use that information. Under the alignment sub-dimension, a teacher shifting to higher levels of expertise changes from only considering the assessment task, to considering

what learning objective the assessment task is connected to; to considering the task, learning objective, and evaluative criteria; to ultimately considering all components of the assessment triangle in addition to literature informing the assessment design. However, the purpose of the rubric is to interpret expertise in terms of teachers *considering* elements of assessment, and not interpreting the actual alignment of learning objective, task, and criteria. Interpreting the alignment among vertices of the assessment triangle involves a deeper level of analysis beyond the scope of the rubric (see Lyon, 2011). For instance, a teacher may discuss specific evaluative criteria that are not matched with the actual learning objective.

Assessment task exemplars. At a level one expertise, the teacher either does not consider an explicit assessment task *or* describes generic assessment forms. For instance, in his/her assessment plan, the teacher might discuss the types of questions to ask:

> Assess them by: Them knowing the differences of Acid/Bases on pH meter. How it affects a system: Buffer system. How it affects solubility of solution. Do they know the formula/equation? (Darlene, Survey 1 assessment plan)

In this example, Darlene plans the assessment around the *content* being assessed, not connecting content with a particular assessment *form*, such as through oral questions, multiple-choice questions, written constructed-response questions, etc.

At level two, the teacher considers a specific assessment form, but there is no indication that the assessment can do more than assess rote, factual knowledge:

> I would have a written test asking questions…such as does an acid or base produce an H+ ion in aqueous solutions. (Hallie, Survey 1 assessment plan)

Although Hallie considers a written test, the assessment task requires essentially a yes or no answer, instead of eliciting how the student applies, connects, reasons about knowledge (i.e., thinks) to arrive at the answer.

At level three, the teacher considers at least one assessment task that elicits some scientific thinking, such as laboratory reports, presentations, problem-solving tasks, and probing oral questions:

> I would work with students in small groups and ask them to explain their observations, as well as questions such as: is the image inverted or upright; is the image larger or smaller than the object, is it real or upright; could you summarize what happens to the image as the object gets closer to the lense [*sic*]. These Q's [questions] are meant to discover commonalities in S [student] thinking to guide formative assessment, & to lead Ss to consider key ideas. (Dean, Survey 3 assessment plan, underlined original)

Finally, at level four, the assessment task not only elicits scientific thinking, but engages students in authentic scientific practices or discourses that are contextualized within students' home, community, or local environment. Although there are no exemplars to represent this level, such tasks might include having students explore the water quality of a local pond or engaging in a debate surrounding the impact of development around a local pond.

Alignment exemplars. At level one, the teacher might consider what to assess or the assessment task, but makes no explicit connection between the two through a learning objective. In the first exemplar presented, Darlene considers what to assess; however, there is no indication of what she wants students to learn.

At level two, the teacher does indicate that the assessment task is connected to some learning objective (specific or general) or considers generic indicators of meeting the learning objectives:

> I will give a lab on acid and base and have them [students] complete a worksheet [to see] if they understand the definition of Acid/Base. (Darlene, Survey 2 assessment plan)

Darlene has now considered the connection between the assessment tasks (lab/worksheet) and the objective (Acid/Base definition). The connection between assessment task and learning objective is consistent with an alignment between the cognition-observation vertices of the assessment triangle as well as Wiggins and McTighe's (2005) understanding by design model.

At level three, the teacher considers both the alignment between assessment task and learning objective and the alignment to the evaluative criteria:

> The main assessment for this lesson sequence is the lab report. This lab report will provide the information for me to see if students understand how to identify metals, nonmetals and metalloids and their properties thus telling me if they have met the lesson objectives. The feedback I am focusing on providing is on writing a sound purpose in the lab report and how to begin thinking about error in the lab. (Darlene, PACT commentary)

Darlene has considered the learning objective (how to identify metals, nonmetals and metalloids and their properties) in relation to the assessment task (lab report) as well as general evaluative criteria (writing a sound purpose and error in the lab). Although the criteria might not be in fact aligned with the learning objective (again, this requires a deeper analysis), she is considering all three vertices of the assessment triangle model.

Finally, at level four, not only is there a connection among learning objective, task, and criteria, but the teacher also makes explicit mention of research or theory that informs the selection or enactment of particular assessment tasks:

> According to the SIOP model, there are 8 key concepts for making content and language accessible for not only English Language Learners, but also all students. I will be using these key concepts to adapt my lessons to make content more clear to students, and also to help with academic language acquisition. The first concept is preparation, so I will spend one day reviewing concepts that students had learned in previous units. I will have students look at diagrams they had drawn for the rock cycle, so they can visually see the diagram that we will be going over so they are fully prepared for the lab. (Glenda, PACT commentary)

Glenda designed instruction (reviewing concepts and looking at diagrams) as part of the assessment task (the lab), based upon literature related to EL learning, thus demonstrating a connection between theory and assessment task.

Use Dimension

The Use dimension generally represents the shift from assessment as strictly an evaluative tool to assessment as an instructional tool, and consists of the *curricular context* and *cycle of inquiry* sub-dimensions. Although assessment can be used as an instructional tool in the moment-to-moment interactions of classroom teaching (e.g., asking probing questions), the focus of this dimension is on using assessment embedded in the curriculum that can be planned, collected, and analyzed. The curricular context sub-dimension represents the extent to which the teacher considers how the assessment fits into the curricular context via the *placement* and the *purpose* of assessment. As teachers demonstrate higher levels of expertise, they shift from only describing the assessment to describing the assessment in relation to instruction,

to describing the specific placement and purpose of assessment, to describing multiple assessments that build on each other to support science learning.

The cycle of inquiry sub-dimension represents the extent to which the teacher considers assessment strategies that can support students' science learning. As teachers demonstrate higher levels of expertise, they shift from not considering formative assessment strategies, to recognizing the need for formative assessment, to considering specific formative assessment strategies. At the highest levels, they organize their assessment plan around a complete cycle of inquiry, consistent with models of formative assessment (Bell and Cowie, 2001; Harlen, 2003; McMillan, 2007; Ruiz-Primo & Furtak, 2007).

Curricular context exemplars. At level one, the teacher only considers the assessment and not when assessment will occur in the curricular unit or the purpose of the assessment, such as in Darlene's Survey 1 assessment plan.

At level two, the teacher considers assessment as part of the unit, but is not specific about the placement or purpose of assessment:

> At the end of the unit, I will assess with a couple scenarios and ask them to find whether a trait is dominant, recessive, codominant or incomplete dom[inant]. (Whitney, Survey 2 assessment plan)

In the preceding example, Whitney does mention the placement (end of the unit), but does not consider why they are assessing at the end of the unit, unlike the following example of a level three response:

> The first thing I would do is have students write down everything they know about earthquakes so I can see where I need to start teaching based on their prior knowledge. (Hallie, Survey 2 assessment plan)

Hallie clearly describes the placement of assessment (the first thing I would do) and assessment's purpose (elicit prior knowledge). Thus, assessment serves a purpose other than just evaluating student learning. It can be used throughout the unit as information to guide instruction.

Finally, at level four, the teacher considers the placement and purpose of multiple assessments throughout the curricular unit that can build on each other (i.e., it is clear why one assessment follows another):

> In a unit on convex & concave lenses, I would assess S's learning in multiple ways. I would start with a lab where Ss find the image of an object refracted through convex & concave lesnes [*sic*]. I would work with students in small groups and ask them to explain their observations, as well as questions such as: is the image inverted or upright; is the image larger or smaller than the object, is it real or upright; could you summarize what happens to the image as the object gets closer to the lense [*sic*]. These Q's are meant to discover commonalities in S thinking to guide formative assessment, & to lead Ss to consider key ideas. I would follow up the lab with a think-pair-share based on the previous questions. The "share" is an informal yet formative assessment in that I get to hear how Ss think about the key ideas, and at the same time scaffold content and language through my responses. At this point, a quick quiz where Ss respond by completing a diagram could be administered to check for understanding in a more visual and less language-dependent fashion. (Dean, Survey 3 assessment plan, underlined original)

Dean goes through a natural assessment progression – beginning the unit with a lab to elicit commonalities in student thinking (prior knowledge), a think-pair-share building from the prior knowledge previously elicited, and then a quiz. Overall, there is not just a purpose for a single assessment but also a purpose for the sequence of assessments.

Cycle of inquiry exemplars. At level one, the teacher does not consider assessment for a formative use and thus does not consider providing feedback,

modifying teaching/instruction based upon assessment information, nor does the teacher involve students in self-assessment and other self-regulated learning strategies.

At level two, the teacher considers, but is not specific about formative assessment strategies:

> If students are working along or with partners, I can have them turn in work to get feedback on. (Teresa, Survey 3 assessment plan)

Although Teresa considers feedback, she is not specific about what type of feedback or what kind of feedback, such as Dean, who demonstrates a level three expertise:

> I think comments written next to students' responses would be best; things like: "interesting, what about…" (Dean, Survey 2 assessment plan)

Finally, at level four, the teacher considers an entire inquiry cycle, including the learning objective, what patterns or alternative conceptions to look for while assessing, and what steps (feedback/instruction modification) will facilitate student attainment of the learning objective. The prior exemplar from Dean's third assessment plan represents a level four in the cycle of inquiry sub-dimension as well. Although Dean does not have a focused learning objective, he has identified the general topic of the unit (convex & concave lenses). He considers using questions during group work to looks for patterns in student thinking, uses a think-pair-share as a next step that is informed by student responses to the questions he asks, and once again looks for patterns in student thinking – suggesting a specific recipe for future action (Wiliam & Leahy, 2007).

Equity Dimension

The equity dimension consists of the *fairness* and *access* sub-dimensions. The fairness sub-dimension represents the extent to which teachers incorporate strategies that make assessment fairer for ELs (i.e., ELs have a more appropriate opportunity to demonstrate scientific knowledge and thinking). Increased expertise is reflected by a shift from *no awareness* of sociocultural influences on assessment to *recognition* of sociocultural influences to strategies that *either modify language or support the influence of language/culture*. At the highest levels of expertise, teachers would draw on and *incorporate* students' culture and language (Solano-Flores & Nelson-Barber, 2001).

The access sub-dimension represents the extent to which the teacher considers how to use assessment to provide ELs with continued access to a rigorous science curriculum. In particular, the criteria for the access sub-dimension examine the presence of scientific discourse, opportunities for language development, and equitable participation during the assessment. Teachers demonstrating higher levels of expertise change from not using assessment to promote complex scientific thinking or language development, to assessment that does promote complex scientific thinking, to specifying how the assessment helps ELs in particular access science curriculum through full participation, complex scientific thinking, or language development. At the highest level, the teacher would also provide feedback tailored to ELs.

Fairness exemplars. At level one, the teacher does not consider the fairness of assessment for ELs, while at level two the teacher may consider that (a) students

come in with various backgrounds, (b) language/culture influence assessment performance, (c) multiple forms of assessment should be used, (d) evaluative criteria should be set, or (e) that assessment features (content, structure) should match the context of instruction. The following examples represent various ways to demonstrate level two expertise:

> summative assessment unit test w/ multiple choice, short answer, show work & explain. (Yvonne, Survey 3 assessment plan) *[multiple forms of assessment]*

Although ELs might not be given a fair chance to demonstrate what they know on a single assessment, using multiple forms of assessments might elicit a wider range of knowledge (Wolf, Bixby, Glen III, & Gardner, 1991), which is important for two reasons. First, ELs might better demonstrate what they know through a different modality. Second, teachers can elicit a variety of scientific knowledge that one assessment may not fully elicit. This corresponds to having an increased repertoire of assessment tasks.

> For presentations, students would be supplied rubrics so that they know what is expected of them. (Matt, Survey 2 assessment plan) *[set evaluative criteria]*

By providing a rubric or other form of evaluative criteria to students, ELs have expectations set for them, which can also serve as another linguistic resource to draw on while engaging in the assessment.

At level three, the teacher considers at least one specific strategy that draws attention to the influence of language/culture on assessment, such as modifying the assessment, scaffolding language, or differentiating assessments:

> Students would do a presentation and a final exam and…decide which could weigh more on their grade. (Michael, Survey 1 assessment plan)

> I believe I have made these concepts accessible in a few different ways. Firstly, the glossary is a really important way to support students who may have a language barrier or who may need some repetition or extra time to fully let a term soak in. … When they work in groups, I encourage student equity buy [*sic*] persuading groups to 'look out' for all their members, because they don't know who I am going to call on for the answer. This can include students who may have a language barrier and can also foster a sense of solidarity with their peers (Barton, 2009). Included in this activity are visuals, semantic webbing, and opportunities for different learning styles, which all support and scaffold the varying needs of my students. (Whitney, *Learning Theories* final project)

Finally, at level four, the teacher does not approach culture/language as a construct interfering with content, but rather as a resource that can help contextualize assessment and curtail linguistic challenges:

> The linguistic demands for this assessment are the same as those needed for the activities. For ELL [English language learner] students, the pictures of the flowers and the pollinators will be helpful, but the written descriptions of flower and pollinator characteristics could be better supportive by also being provided in their native language. An example of matching a pollinator to its flower could also be included to support ELLs. (Lauren, *Learning Theories* Final Project)

Lauren describes two strategies – visuals and using ELs' native language. By using ELs' native language, she is not assuming that simple language modification will make assessment fairer, instead she is using language as a resource to support ELs, consistent with the notion of cultural validity (Solano-Flores & Nelson-Barber, 2001).

Access exemplar. At level one, the teacher does not consider assessment that allows students to access scientific discourse either through reading, writing, or talking, while at level two the there is a presence of scientific discourse in the assessment:

> I will work in student groups and ask them to explain their observations. (Dean, Survey 3 assessment plan)

Dean considers an assessment that allows the students to engage in scientific discourse by explaining observations. However, Dean does not explicitly state how engaging in scientific discourse through assessment can benefit students. Thus, a distinction between levels two and three is considering scientific discourse versus explicitly connecting how the assessment provides students with access to a rigorous science curriculum and/or develop language:

> Throughout all activities I will be discussing concepts in any and all ways that promote understanding for my students. If a student describes a concept or scientific idea in a way that makes sense to them, I will use that example and repeat its meaning using scientific language. The use of concept maps and charts, in which students can see multiple ways of explaining the same idea, in both conversational and scientific language, will promote and increase students' abilities to participate in scientific conversations in both oral and written form. (Whitney, PACT commentary)

Finally, at level four, not only does the teacher explicitly consider how assessment can help ELs fully participate in science, promote scientific discourse, or develop language, but he or she also consider how to tailor feedback in these areas to ELs. While I did not observe any instances of level four expertise, the exemplars representative of level three could become level four if, for instance, Dean or Whitney also discuss how they would provide ELs with feedback on vocabulary use while they explain their observations or make concept maps.

Putting it All Together: Hallie's ***Learning Theories*** **Final Project**

Previously, teacher responses were used to exemplify the various levels within *each* sub-dimension of the rubric. To demonstrate what is learned by analyzing a

single response through *all* sub-dimensions, I present an excerpt from Hallie's *Learning Theories* final project. In the excerpt, Hallie describes how she would assess the topic, *Galaxies, the Sun, and Stars*. Table 2 displays her scores on each sub-dimension.

Table 2

Assessment Expertise Scores for Hallie's Learning Course Final Project

	Construction		Use		Equity	
	Assessment task	Alignment	Curricular context	Cycle of inquiry	Fairness	Access
Score	3	2	2	1	3	2

Note. Possible range of scores: 1-4.

> There are many different ways to assess student learning. One way that I will assess my students is by using a take home work sheet that the students must complete. This worksheet will help teachers understand whether or not their students understood the information that was presented to them. Another way that I will be assessing students is by grading their presentations. This will allow teachers to see whether or not the students learned the information that they are presenting. During the presentations the other classmates will have a worksheet with information to fill out about other group's stars or galaxies. Looking at the assessment now, I realize that adding pictures to the worksheet would greatly benefit ELL and LEP [limited English proficient] students. Another idea to keep in mind while assessing groups is making sure to not grading [*sic*] strictly on presentation skills. Some ELL and LEP students may not feel comfortable talking in front of the class. Asking students to write down what parts of the group work they helped on would make for a fairer grade.
>
> -Excerpt from Hallie's *Learning Theories* final project

Regarding the Use dimension, Hallie describes how the general purpose of one assessment – the purpose of the take home worksheet – "will help teachers understand whether or not their students understood the information that was presented to them." This purpose is not consistent with a formative use of assessment.

Furthermore, she does not indicate the sequence of assessment, or when in the unit assessment occurs. Thus, Hallie has demonstrated level two expertise on the curricular context sub-dimension. Hallie does not consider assessment strategies such as providing feedback or modifying teaching/learning to support students' science learning, thus demonstrating a level one expertise on the cycle of inquiry sub-dimension. If researchers examine Hallie's assessment expertise from merely a formative lens, they might conclude that Hallie has limited expertise in classroom assessment. However, by analyzing her response through multiple dimensions, we can capture a much more complete picture of her assessment expertise.

Regarding the Construction dimension, Hallie considers particular assessment tasks. Although some assessment tasks are generic (e.g., worksheets), she has described a project presentation that allows her to elicit student thinking, thus demonstrating a level three expertise for the assessment task dimension. She understands that the tasks allow teachers "to see whether or not the students learned the information," but she does not articulate exactly what the learning goal is or how she will grade (e.g., a rubric? a key? a scoring system?). Thus, she has considered generally the connection between the cognition and observation vertex of the assessment triangle model, demonstrating a level two expertise in the alignment sub-dimension, but not enough for a level three.

Finally, Hallie does consider ways in which the assessment might be unfair for ELs. Specifically, she acknowledges that limited English proficiency might have an effect on student performance while talking in front of the class. She goes beyond

this recognition to suggest practices that draw attention to the influence of language. In terms of the presentation, she suggests a written component (using another assessment task) and adding pictures to the worksheet (a language support). Thus, she has demonstrated a level three expertise in the fairness sub-dimension. Through the presentations, ELs also might access a rigorous science curriculum by being given opportunities to talk science, although Hallie does not explicate the exact nature of the presentation and thus does not show how engaging in the presentation would allow ELs to talk science. Hallie has demonstrated a level two expertise in the access sub-dimension.

In summary, written products from a teacher can be analyzed via the rubric and used to infer the teacher's level of assessment expertise in the various dimensions of the CUE Assessment Expertise Framework. More importantly, by incorporating multiple dimensions, researchers have a better representation of teachers' assessment expertise.

Cautions and Promises for Studying Science Teachers' Assessment Expertise

The CUE Assessment Expertise Framework is grounded in sociocultural theory and builds on literature from science education, educational assessment, and effective teaching for ELs. The framework consists of the following conceptual dimensions: (a) constructing theoretically cohesive assessments that elicit student thinking, (b) using assessment to support students' science learning, and (c) assessing in ways that are equitable for ELs. Exemplars were presented to illustrate how the framework translated into a rubric used to capture various aspects of expertise. In the

final exemplar, Hallie demonstrated levels one/two expertise in the Use sub-dimensions, yet demonstrated levels two/three expertise in the Construction and Equity sub-dimensions on her *Learning Course* final project. Thus, each dimension shows different aspects of a teachers' assessment expertise, which many not be at the same level across dimensions. Through the description of application of the framework, this paper has argued for a multidimensional approach to provide a more complete and detailed account of science teachers' assessment expertise, given the complexity of science classroom assessment.

While the CUE Assessment Expertise Framework and Rubric were useful in the investigation, I foresee future work to refine how science education researchers study science teachers' assessment expertise. An important aspect of my study was to evaluate teacher cognition – how teachers think and the decisions they make – that can be difficult given the ambiguity of constructs, the inability to assess directly such constructs, and the time-consuming nature of analytical methods (Kagan, 1990). The CUE Assessment Expertise Rubric assuaged such difficulties by translating the conceptual dimensions of assessment expertise into well-defined observable phenomena that could draw inferences about teachers' assessment planning. Evident by the exemplars and through various lenses, the CUE Assessment Expertise Rubric functioned well enough to discern the range of expertise of 11 teachers in one teacher education program. However, the rubric has not been sufficiently validated for use in large-scale study. Therefore, next steps may include validating instruments, protocols, and scoring rubrics with a larger data set to aid in intervention research. Alternatively,

researchers can use the rubric to compare assessment expertise between different groups of teachers (e.g., beginning versus experienced). Furthermore, researchers can identify "expert" teachers and closely study how they assess science learning, thus providing more insight into assessment expertise. This approach is consistent with teaching expert research in which an "expert prototype" that demonstrates teaching knowledge, efficiency, and insight serves as a model for expertise (Sternberg & Horvatch, 1995).

There are other important considerations while analyzing science teachers' assessment expertise, such as the unit of analysis and the context of the data source. The exemplars represented various sources of text and various text lengths (excerpts were presented in some cases and entire responses in others). Should researchers analyze an entire response to a single prompt? How do analytical tools, such as rubrics, hold up when triangulated with other data sources? Triangulation can inform the construct validity of the rubric and whether other instruments are needed to tap into the complexities associated with assessment expertise. For example, I drew on the rubric to refine the coding scheme used for qualitative analyses (e.g., different conceptualizations of equitable assessment), thus providing stronger support for growth in assessment expertise. Furthermore, can researchers apply the rubric to completely different contexts, such as observation of assessment practices? In the context of my study, I analyzed the content of independent products, such as the entire response to the survey assessment planning prompt and the *Learning Theories* final project. I used the same rubric for both data sources; however, it took time to

train scorers on how to apply the rubric across both data sources. It will be beneficial for researchers to use a single (yet tailored) rubric to analyze different data sources in a consistent manner, all while considering the unit of analysis.

Another caution is that I did not design the rubric to imply a linear development from levels one to four. What the rubric *does* do is help infer shifts in expertise by both looking across participants and looking across various dimensions, which can answer a range of questions. For example, does an individual's expertise at using assessment formatively seem to change significantly from the beginning to the end of the teacher education program, while their expertise in other dimensions do not? Furthermore, qualitatively, what is different about changes across the dimensions?

The value of this paper lies its use of a conceptual framework and analytical tool (i.e., rubric) to study science assessment from a teacher-centered perspective. Once researchers develop, validate, and pilot the next generation of science assessments, which elicit scientific practices and potentially utilize technological innovations, it becomes important to understand how science teachers use assessments in conjunction with other assessments, curriculum, and instruction to support science learning for all students. Furthermore, if science teachers have not developed assessment expertise fully, then simply presenting them with validated assessment will not ensure that the teachers will use the assessments in ways that are supportive and equitable.

If researchers aim to expand upon case studies that examine focused aspects of assessment (cf., Treagust, Jacobowitz, Gallagher, & Parker, 2001), we need to unpack the complexities of science classroom assessment, which the CUE framework and Rubric attempt to do. By examining multiple data sources through all three dimensions, researchers can provide a more complete and nuanced analysis of assessment expertise. Researchers can address what science teachers are doing with assessment strategies while assessing specific scientific content (Coffey, Hammer, Levin, & Grant, 2011) and, the barriers they may face, and even find relationships between expertise and student achievement in science. Furthermore, researchers can study *how* teachers become prepared to assess in ways that align with science education reform and incorporate such assessment practices in their teaching. While the science assessment literature has focused on what teachers should know and be able to do with assessment, researchers have given little attention to the process of getting there.

Finally, the framework can inform instructional focus for teacher preparation and professional development. No longer can preservice science teachers get by with a cursory treatment of assessment in which they learn how to replicate standardized tests (Shepard, 2006). Teachers must (a) be equipped with a repertoire of assessment forms that elicit various kinds of knowledge and skills, (b) have knowledge of how to use assessment to support learning, and (c) hold particular views of learning and assessment values (Abell & Siegel, 2011). Furthermore, given that teachers are likely to be placed in culturally and linguistically diverse classrooms, they must also

consider the influence of language and culture on the assessment process as well as other issues of equity. Many assessment practices seem tacit to teachers (Bell & Cowie, 2001), and teachers would benefit if researchers better clarify the constructs and translate them into observable practices. As earlier described, for my exploratory study, I developed a series of activities so that preservice science teachers can learn about principles associated with each dimension, reflect on strategies to carry out the various dimensions of expertise, and discuss their views on each dimension with their peers. Furthermore, the rubric can help teacher educators consider a learning progression of assessment expertise. For instance, in the year following my data collection, the rubric informed my own teaching. While providing assessment-related instruction, I first helped a cohort of teachers recognize that assessment tasks include more than tests and exams – that they share a common goal of generally finding out what students know and can do. Then, I provided greater focus on the alignment between assessment task and learning objectives. After this, I introduced the general idea that specific criteria should connect to the learning objective and task. The important point here is that assessment is complex and to adequately prepare teachers, teacher educators need to carefully scaffold instruction.

In conclusion, studying science teachers' assessment, including how teachers develop expertise and enact it in their classroom, is important for research in science assessment. However, studying the teachers, and not the assessment itself, involves a different approach, both theoretically and methodologically. The CUE Assessment Framework and Rubric are two such tools that can contribute to this continued

conversation. Assessment in classroom contexts is complex and of great importance to how students learn in academic settings, such as science classrooms. If science education is to embrace an assessment culture, then researchers must understand all aspects of science classroom assessment and continue unpacking its complexity so that science teachers can ultimately be prepared to assess in ways consistent with science education reform and responsive for diverse learners.

References

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1-28.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer Netherlands.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education.* Dordrecht: Kluwer Academic Publishers.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Buckingham: Open Univ Press.

Black, P., & Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Bransford, J., Brown, A.L. & Cocking, R. (2000). *How people learn.* Washington, DC: National Academy Press.

Broadfoot, P. (1996). *Education, assessment, and society: A sociological analysis.* Buckingham, PA: Open University.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*(10), 1109-1136.

Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York, NY: The Free Press.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., & Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment, 11*(3&4), 237-263.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* New York, NY: Routledge.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*(1), 355-392.

Gipps, C. V., & Murphy, P. (1994). *A fair test?: Assessment, achievement and equity.* England: Open University Press.

Glaser, R., & Chi, M. T. (1988). Overview. In M. Chi, R. Glaser & M. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Erlbaum.

Haertel, E., Moss, P., Pullin, D., & Gee, J. P. (2008). Introduction. In P. Moss, D. Pullin, J. P. Gee, E. Haertel & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn*. England: Cambridge University Press.

Harlen, W. (2003). *Enhancing inquiry through formative assessment*. San Francisco, CA: Exploratorium.

Hill, M., Cowie, B., Gilmore, A., & Smith, L. F. (2010). Preparing assessment-capable teachers: What should preservice teachers know and be able to do?. *Assessment Matters, 2*, 43-64.

Kagan, D. M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks principle. *Review of Educational Research, 60*(3), 419-469.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English language. *Educational Researcher, 27*(4), 12-21.

Luft, J. A. (1999). Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education, 10*(2), 107-121.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

Lyon, E. G. (2011). Beliefs, practices, and reflection: Exploring a science teacher's classroom assessment through the assessment triangle model. *Journal of Science Teacher Education, 22*(5), 417-435.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. *Examining Pedagogical Content Knowledge, 6*(2), 95-132.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

McMillan, J. H. (2007). *Formative classroom assessment: Theory into practice.* New York, NY: Teachers College, Colombia University.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future*. London: King's College London.

National Center for Education Statistics (2009). *Science framework for the 2009 national assessments of educational progress*. Author.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

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

National Research Council. (2001). The nature of assessment and reason from evidence. In J. W. Pellegrino, N. Chudowsky, & R. Glaser, R. (Eds.), *Knowing what students know: The science and design of educational assessment* (pp. 37-54). Washington, DC: National Academy Press.

Oakes, J. (1990). Multiplying Inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science (Report No. NSF-R-3928. Santa Monica, CA: Rand.

Organisation for Economic Co-operation and Development. (2009). Pisa 2009 science framework. In *Pisa 2009 assessment framework – Key competencies in reading, mathematics, and science* (pp. 125-148). Author.

Pease-Alvarez, L., & Hakuta, K. (1992). Enriching our views of bilingualism and bilingual education. *Educational Researcher, 21*(2),4-19.

Popham, W. J. (2006). *Assessment for educational leaders.* Boston, MA*:* Allyn & Bacon.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295-314.

Shaw, J. M., Bunch, G. C., & Geaney, E. R. (2010). Analyzing language demands facing English learners on science performance assessments: The Sald framework. *Journal of Research in Science Teaching, 47*(8), 909-928.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4). Westport, CT: Praeger Pub Text.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching, 44*(6), 864-881.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-573.

Sternberg, R. J., & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24*(6), 9-17.

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

Tamir, P. (1998). Assessment and evaluation in science education: Opportunities to learn and outcomes. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (Vol. 2). Dordrecht: Kluwer Academic Publishers.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context*. England: Cambridge University Press.

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. *Science Education, 85*(2), 137-157.

Valdes, G. (2001). *Learning and not learning English: Latino students in American schools*. New York, NY: Teachers College Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: The MIT Press.

Webb, N. (2002). Assessment literacy in a standards-based urban education setting. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design. Alexandria, VA:* Association for Supervision and Curriculum Development.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative classroom assessment* (Vol. 22, pp. 29-42). New York, NY: Teachers College Press.

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*(2), 31-37.

Appendix
CUE Assessment Expertise Rubric

Level 1	Level 2	Level 3	Level 4
Construction: *Assessment task*			
Does not consider an explicit assessment task (describes generic forms (e.g., tests, quizzes, ask questions)	Considers a specific assessment task, but the task is **not appropriate** for eliciting student thinking (e.g., asks for definitions, factual information)	Considers at least 1 assessment task that elicits **some** student thinking (i.e., tasks are open-ended and do not have a correct/incorrect response)	Considers at least 1 assessment task that elicits **and engages** students in **authentic scientific practices** (guided/open inquiry; research situated within a real life context)
Construction: *Alignment*			
Considers what to assess **AND/OR** the assessment task, but does not align the two	Considers the alignment between the assessment task and the learning objective (objective may be general) **OR** Considers the general use of evaluative criteria (e.g., a rubric or broad indicators of mastery)	Considers the alignment among the assessment task, the **specific** learning objective, and **specific** criteria (must do more than just mention a rubric)	Considers the alignment among the assessment task, the **specific** learning objective, and **specific** criteria (must do more than just mention a key or rubric) **AND** considers some theory that informs the assessment design
Use: *Curricular context*			
Does not consider the curricular context (placement or purpose of assessment in the curriculum)	Considers assessment as part of the curricular unit, but is **not specific** about sequence of assessment or the purpose of assessment (may include a "summative test" if they also discuss some aspect of the curricular unit)	Considers the placement and purpose of **at least one** assessment (e.g., KWL at the start to activate prior knowledge). Must have more than a "summative test"	Considers the placement and purpose of **multiple assessments** throughout the unit that can build on each other and collectively contribute to student learning
Use: *Cycle of inquiry*			
Does not consider assessment strategies to support science learning (e.g.,	Considers (but is not specific about) assessment strategies that may support science learning, such as (a) giving feedback, (b)	Considers **specific** assessment strategies that may support science learning	Considers an entire cycle of inquiry (what do I want students to learn? → what patterns/alternative

feedback, modifying instruction)	modifying instruction, or (c) actively involving students in the assessment process (e.g., assessment design, self-assessment) **OR** just mentions "formative assessment"		conceptions will I look for? → what steps (feedback/modify instruction) will I use to help student attain the learning objective?
Equity: *Fairness*			
Does not consider the fairness (bias) of assessment for ELs	Considers the fairness (bias) of assessment for ELs, such as that (a) students come in with various backgrounds, (b) language /culture influence assessment performance, (c) multiple forms of assessment should be used, or (d) assessment features (content, structure) should be match to the context of instruction	Considers at least 1 strategy that ***draws attention*** to the influence of language and culture on assessments (e.g., modify language, scaffold language, modeling, differentiate assessments)	Considers at least 1 strategy for ***incorporating*** students' language/ culture in the design/use of assessment
Equity: *Access*			
Does not consider opportunities for ELs to engage in complex thinking, develop language, or fully participate	Considers assessments that allows students to talk science, read or write authentic science texts, or learn the language of science but does not explicitly link this to a need for ELs	Considers (explicitly) how assessment can help ELs fully participate in science, promote complex thinking, or develop language	Considers (explicitly) how assessment can help ELs fully participate in science, promote complex thinking, or develop language **AND** how to provide feedback tailored to EL needs

CHAPTER 2

ROUGH WATERS AND SMOOTH SAILING: CHARTING THE CHANGES OF SECONDARY SCIENCE PRESERVICE TEACHERS' ASSESSMENT EXPERTISE

Abstract

Through a mixed methods analysis, this paper reports on the ways in which the assessment expertise of 11 secondary science preservice teachers (the "teachers") changed as they participated in a teacher education program where the author had led assessment-focused instruction in three courses. Assessment expertise was captured through surveys, interviews, and teacher products collected throughout the program. Broadly, the findings indicate that the teachers changed in both their understanding of and facility with using assessment formatively. In depth descriptive and qualitative analysis revealed more subtle aspects of change. The teachers (a) expanded their assessment task repertoire, (b) considered the alignment of assessment tasks with learning objectives, (c) shifted from an evaluative to instructional view of assessment, and (d) became aware of the role of language while assessing. Conversely, some aspects of assessment expertise did not change, such as considering fully the alignment of learning objectives to evaluative criteria. By identifying and describing "rough waters" and "smooth sailing" for the teachers, this paper suggests potential barriers while being prepared to assess science learning, warranting further exploration.

Introduction

Assessment plays a critical role in secondary (middle school and high school) science classrooms, both in reporting what students know, which affects their advancement throughout the curriculum, and in supporting students' science learning. Over the last two decades, science education research has led to new forms of assessment (e.g., performance assessment) that elicit student thinking as opposed to factual recall. Furthermore, scholars have theorized about how teachers can use assessment to inform science instruction while mounting empirical evidence that using assessment formatively raises student achievement (Black & Wiliam, 1998a; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Wilson & Sloane, 2000). The confluence of new assessment forms and functions leaves individuals preparing to become secondary science teachers with a daunting task, one that may take considerable effort to accomplish as they navigate through teacher preparation programs. Yet, these programs often do not adequately prepare teachers to assess in ways consistent with emerging views of learning and assessment (Shepard, 2006; Stiggins, 2002). Adding to the difficulty, preservice science teachers are entering classrooms with an increasingly diverse student population, in terms of students' native language and English language proficiency (National Center for Educational Statistics, 2011).

Despite advances in science assessment design and use, there exists but a handful of empirical studies that describe preservice science teachers' assessment knowledge, beliefs, and application of such understandings. It is even scarcer to find

studies that examine changes across an *entire* teacher education program. Furthermore, to my knowledge, no studies analyze those changes while attending to issues of equitably assessing English learners (ELs). Given the importance of assessment, yet dearth of literature on preservice teachers' capability to assess, this study explores in what ways 11 secondary science preservice teachers' (hereon referred to as the "teachers") understandings of and facility with assessment (i.e., their assessment expertise), changed while participating in a teacher education program. The study's goal was to build upon other small-scale, descriptive studies of science teachers' assessment expertise by exploring a new context (preservice science teachers exposed to theoretically guided assessment instruction focused on linguistically diverse students) and longitudinally analyzing assessment expertise through *multiple* conceptual dimensions.

Science Teachers' Assessment Expertise: A Brief Review of the Literature

Researchers over the last decade have used the term assessment literacy to describe what science teachers need to know and be able to do with assessment (Abell & Siegel, 2010; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Siegel & Wissehr, 2011; Webb, 2002). While implementing and studying assessment-related professional development for science teachers, Gearhart et al. (2006) focused on the "*relationship* between understandings of assessment concepts and facility with assessment," (p. 259, emphasis added) a relationship they refer to as assessment *expertise*. I also employ the term assessment expertise, as opposed to literacy, given the tradition in teacher cognition/learning studies to study expertise (Clark &

Peterson, 1986; Schemp, Tan Manross, & Fincher, 1998; Shulman, 1986) and the well grounded theoretical basis of what it means to become an expert (Bransford, Brown, & Cocking, 2000; Dreyfus & Dreyfus, 1986). Researchers have explored various aspects of teachers' assessment expertise in a range of contexts, such as self-reported understanding, responses to course assignments, and observation of classroom practice. Each context adds to the knowledge base in unique ways.

Assessment Understanding: Beliefs and Knowledge

Teacher beliefs about learning and assessment values are critical while preparing assessment-capable teachers (Abell and Siegel, 2011; Black & Wiliam, 1998b; Hill, Cowie, Gilmore, & Smith, 2010). However, studying beliefs is problematic given the various construct definitions (Jones & Carter, 2007; Pajares, 1992), such as whether belief is synonymous with knowledge, divorced from it, or intricately related to it (Smith and Siegel, 2004). In this study, beliefs are conceptualized as personal constructs, indicated by propositions considered to be true by the individual, because they are based on *personal* judgment and evaluation (Luft, Roehrig, Brooks, & Austin, 2003). Knowledge is intricately linked to beliefs and best described as socially constructed propositions based upon some evidential support. Furthermore, both knowledge and beliefs can either be "professed," – articulated by the teachers themselves through interviews or other written/oral responses – or "attributed" – identified by the researcher during analysis (Aguirre & Speer, 2000).[1]

The literature related to teachers' assessment understanding is sparse and limited in scope. Instead, scholars have theorized more about what teachers *should*

know and believe about assessment then what they *actually* know and believe. Some evidence exists that science teachers' beliefs about science learning (Czerniak & Lumpe, 1996), science assessment (Lyon, 2011), and even fairness (Yung, 2008) influence their assessment practices. It is difficult for interventions and professional support to change practicing teachers' beliefs about science teaching (Lee, 2004). However, there may be more promise for preservice teachers.

Individuals come into teacher education programs with assessment understanding limited to their experiences as students, often associating assessment with "tests," and do not consider the importance of learning goals nor how assessment can be used formatively (Graham, 2005). By the end of teacher education programs, preservice teachers demonstrate some understanding through concern about designing "good" learning objectives, designing assessments that draw valid inferences, interpreting work fairly, and having time to use assessment formatively (Graham, 2005). Regarding formative assessment strategies, Marshall and Drummond (2006) argue that *exposure* to basic principles is not enough to enact changes in assessment practices. Content-specific teacher preparation courses, such as a science methods course, that focus on formative assessment principles and that promote reflection of learning and assessment beliefs have been shown to increase teacher confidence at using formative assessment strategies (Yilmaz-Tuman, 2008). Reflecting on and interrogating assessment beliefs are particularly important given that preservice teachers hold complex conceptions of assessment – beyond just summative versus formative (Remesal, 2011).

Assessment Facility: Applying Assessment Beliefs and Knowledge

More research has studied how teachers apply their assessment knowledge and beliefs. Studies have collected teacher responses to hypothetical prompts, have collected teacher lesson plans, and have observed teachers' during day-to-day teaching. Even with adequate knowledge of basic assessment principles and current assessment tenets, preservice teachers do not apply fully their knowledge while engaging in assignments associated with program coursework (Campbell and Evans 2000; Siegel and Wissehr, 2011). Teaching experience and characteristics of the situation (e.g., student and curricular context) are two factors that may influence the decisions teachers make while selecting, in particular, formative assessment tasks (Tomanek, Talanquer, & Novodvorsky, 2008). Thus, while interpreting teacher responses to hypothetical tasks and course assignments, researchers should attend to the particular context – do the teachers know *who* is being assessed or *what content* they are assessing? Although assessment facility may be positively associated with more teaching experience, it is unclear whether assessment facility is noticeably increased by the end of a teacher education program.

Regarding classroom practice, studies are just beginning to articulate how teachers assess science. Research has documented many of the struggles teachers may experience while (a) pre-assessing student learning (Morrison & Lederman, 2003), (b) using assessment information to inform instruction (Otero & Nathan, 2008), (c) aligning assessments to learning goals (Bol & Strange, 1998), and (d) fully aligning assessment tasks and criteria with beliefs about assessment and how students learn

(Lyon, 2011). Similar to analysis of course products, several of these studies indicate that while the teachers may hold assessment understandings consistent with current emphases in science education, they do not productively put these understandings into practice.

Studies have also described successful assessment practices, usually through a formative assessment lens. Drawing on an analysis of formative assessment practices implemented in science classrooms, Ruiz-Primo & Furtak (2007) argued that secondary science students learn more in classes where the teachers engage in such practices. Through a single case study, Treagust, Jacobowitz, Gallagher, & Parker (2001) provided an even richer description of how assessing science formatively looks in classroom practice. Unfortunately, little is known about why some teachers assess in supportive ways while others do not. Among other factors, Aydeniz (2006) has suggested that such barriers may rest on teachers' limited knowledge of assessment in relation to science learning and recommends that teacher educators should help science teachers develop a sophisticated base of content-specific assessment knowledge. Although science teachers may demonstrate some broad knowledge about assessment, they may need support in enriching this knowledge base so that they can more easily apply knowledge to day-to-day science teaching.

Researchers can help science teachers develop content-specific assessment knowledge and, more generally, assessment expertise by focusing on *teacher*-identified assessment practices, building collaboration and trust with the teacher, and recognizing that development takes time (Briscoe & Wells, 2002; Sato and Atkin,

2007). In science methods courses, Buck, Trauth-Nare, and Kaftan (2010) suggest that a combination of case studies, practical fieldwork, and opportunities for self-reflection fosters formative assessment understanding. Researchers have also provided specific support for in-service teachers aimed at helping them analyze and make decisions from student evidence –support that could translate to teacher preparation. Although the specifics of the supports vary, they consistently incorporate tools such as portfolios that aid in changing teachers' assessment knowledge and practices (Darling-Hammond, 2006; Gearhart et al., 2006) and that improve student achievement (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Gerard, Spitulnik, & Linn, 2006; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Sato, Coffey, & Morrthy, 2005; Wilson & Sloane, 2000). Learning about assessment through a variety of carefully designed tasks and experiences can provide scaffolds that support teachers as they transform in their understanding of science teaching and learning (Ash & Levitt, 2003).

In summary, the literature related to science teachers' assessment expertise suggests that under normal circumstances many of the assessment principles currently called for in science education are not being practiced; however, with support teachers can engage in assessment practices in ways that are supportive for student learning. Furthermore, gaps exist in the literature, notably around the assessment of ELs, which if addressed could help us better understand what leads to growth in assessment expertise during teacher preparation.

Research Question

In the context of theoretically guided assessment instruction embedded within three teacher education program courses for a cohort (N = 11) of secondary science preservice teachers, the overarching research question was as followed: *In what ways does the teachers' assessment expertise change over time in the program*?

The overarching question was addressed by analyzing assessment expertise in light of three conceptual dimensions, described in the next section, and by answering the following sub-questions:

1) Do the teachers' beliefs toward current science assessment emphases, expertise levels while planning assessment, or expertise levels while critiquing assessment change significantly from the beginning to the end of the program?

2) What patterns emerge when analyzing the distribution of expertise levels while planning assessment throughout the program?

3) What patterns emerge when analyzing, qualitatively, the teachers' assessment understandings and facility with assessment throughout the program?

Analyzing Assessment Expertise through the Construction-Use-Equity (CUE) Framework

In this next section, I describe the Construction-Use-Equity (CUE) Assessment Expertise Framework. The CUE framework was informed by educational assessment, science education, and EL learning literature and guided both data collection and data analysis. Although not elaborated upon in this paper, the framework was firmly rooted in sociocultural theory[2] (see chapters 1 and 3 for a thorough discussion).

Construction: Constructing Coherently Aligned Assessments

The Construction dimension draws on assessment design principles based on the premise that assessment is a *process of inferring about what students know and can do* (National Research Council [NRC], 2001; Popham, 2006; Shavelson, Ruiz-Primo, Li, & Ayala, 2003). In particular, this dimension uses the assessment triangle (NRC, 2001) as a model to analyze teachers' expertise in this process as well as science education literature to inform *what* science teachers should assess.

The assessment triangle (NCR, 2001) represents how any assessment involves three core components – cognition, observation, and interpretation. The assessment task (observation) used to observe some aspect of learning should draw on theories of how students learn and what is important to learn (cognition). For example, science education frameworks emphasize teaching and assessing scientific thinking and scientific practices, as opposed to recalling a body of scientific knowledge (Millar & Osborne, 1998; NCR 1996, 2012). Furthermore, the criteria used to interpret observations of student learning (interpretation) should also align with the learning objective and theories of how students learn.

Therefore, not only are higher levels of expertise associated with a capacity to solicit scientific argumentation, investigative skills, problem solving, and other forms of complex scientific thinking, but also associated with the teacher's capacity to align the three components of the assessment triangle.

Use: Using Assessment Information to Support Science Learning

The Use dimension draws on the principle that assessment is a critical instructional component that has a considerable effect on student achievement when used in ways that support student learning – in other words, used *formatively* (Black & Wiliam, 1998b). In this study, formative assessment is viewed as an instructional feedback cycle (Bell and Cowie, 2001; Harlen, 2003; McMillan, 2007; Ruiz-Primo & Furtak, 2007), meaning that the teacher iteratively considers "*Where do I want students to get to?*" (i.e., the learning objective), "*What do students now know?*" (by interpreting responses to the assessment task) and finally "*What actions should my students and I take to get them there* (to master the learning objective)*?*"

A teacher demonstrating higher expertise levels in this dimension would consider specific actions to take, such as providing feedback and modifying instruction, which facilitates science learning. Furthermore, at higher levels of expertise, the teacher would have the capacity to embed multiple assessments within the instructional units so that assessments build on each other, as opposed to being used once at the end of a unit. At lower levels of expertise, teachers would only acknowledge generally that they should assess formatively or consider assessment for only evaluative purposes.

Equity: Equitably Assessing English Learners

Language figures prominently in the third dimension, Equity. In any content-area assessment, such as science, a student's proficiency in the language of the assessment is also being evaluated (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on

Measurement in Education [NCME], 1999). Thus, a first step to demonstrating assessment expertise is to recognize how assessment can be biased against ELs due to unfamiliar vocabulary, complex sentence structure, and other language demands (Abedi & Lord, 2001; Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Teachers demonstrate higher levels of expertise by modifying text, accessing multiple modalities (e.g., speaking in addition to writing), and scaffolding language in assessment through contextualization, visual or audio support, sentence frames, etc. (Siegel, 2007).

Equitable assessment involves more than just focusing on assessment fairness (Gipps & Murphy, 1994). It involves understanding how to facilitate both ELs' science learning and English language development (Stoddart, Pinal, Latzke, & Canaday, 2002; Stoddart, Solis, Tolbert, & Bravo, 2010) through assessment. ELs are often relegated to less rigorous science courses due to tracking or perceived ability to engage in complex scientific thinking (Pease-Alvarez and Hakuta, 1992). Therefore, teachers can demonstrate expertise by supporting students' use of scientific discourse (i.e., engaging students in scientific argumentation, explanation) and science vocabulary while reading, writing, and talking in the assessment process. Such exposure in discourse and literacy in science assessment can both provide them access to a rigorous science curriculum and support their science learning.

Research Context

Individuals preparing to become teachers usually do so through a university-affiliated teacher education program. Thus, an important decision in this study was to

select a program that met the following three baseline criteria: (a) I would be able to lead assessment instruction for the teachers, (b) the program would emphasize language and equity issues in education, and (c) preservice teachers would be placed in linguistically diverse classrooms for a teaching practicum.

Hacienda University's Teacher Education Program

The teacher education program selected is affiliated with Hacienda University[3] located in the state of California in the United States. The 12-month long program leads to a California single subject teaching credential and a Masters of Arts in Education. The program groups preservice teachers into cohorts – either multiple subject (grades K-6), secondary English, secondary social studies, secondary math, or secondary science. Beginning in July, secondary preservice teachers take foundational courses in learning and teaching (for diverse students), courses related to teaching in their specific content area (such as science education theory and science methods), and a quarterly seminar course led by teacher supervisors who observe preservice teachers in their teaching practicum. At the onset of program, secondary preservice teachers are assigned a cooperating teacher, whom they first observe, then whose class they begin teaching lessons in, with support from the cooperating teacher. In November, the preservice teachers are assigned a new cooperating teacher, one that matches the desired grade level, and gradually take on more teaching responsibility culminating in complete autonomy of the classroom by the springtime.

The program met all baseline criteria. First, I had prior experience working with faculty in the program. For instance, I had piloted assessment-related instruction

in several courses the previous year, thus allowing me to build a relationship with the faculty, teacher supervisors, and director. Thus, in turn has provided me with the opportunity to access knowledge about the program and facilitate instruction within several courses. Second, the program, evident by its mission statement and program structure, is committed to language and equity issues in K-12 classrooms, which translates to admitting preservice teachers that share in this commitment and placing preservice teachers in culturally and linguistically diverse classrooms, thus meeting the final criterion.

Participants

Eleven (8 women, 3 men, M_{age} = 24 years, age range: 21-26 years) secondary science preservice teachers, the entire secondary science cohort, were invited to participate in this study and each one gave their informed consent to participate. As seen in Table 1, most teachers were self-identified as White non-Hispanic or Asian and native English speakers. Three teachers (Darlene, Matt, and Yvonne) were fluent in a second language. The teachers held degrees in a variety of science-related undergraduate majors that corresponded to the subjects they taught while student teaching. As part of an initial survey, nine of the teachers reported some experience learning about educational assessment – either through their undergraduate education (6 of the 9), observation or teaching in classrooms (4 of the 9), and/or through tutoring (1 of the 9). Some of these specific experiences, reported through an initial interview, included learning how to assess with a variety of forms because of

different learning styles (e.g., kinesthetic) and different orientations (e.g., behaviorist) to assessment, introduction to formative assessment, and standardized testing.

Table 1

Participant Information

Participant	Gender	Race/ Ethnicity	Age	Undergraduate major	Second language proficiency
Darlene	Female	Asian	21	Chemistry	Fluent (Chinese)
Dean	Male	White	26	Physics	Beginning (French)
Glenda	Female	White	21	Environmental Science	None
Hallie	Female	White	25	Health Sciences	Intermediate (Sign language)
Lauren	Female	White	25	Environmental Sciences	None
Matt	Male	Multiracial	25	Biological Sciences	Fluent (Portuguese); Intermediate (Spanish)
Michael	Male	White	22	Physics	None
Teresa	Female	Asian	26	Biological Sciences	Beginning (French, Spanish)
Whitney	Female	White	24	Environmental Sciences	None
Willow	Female	White	21	Biological Sciences	Beginning (Spanish)
Yvonne	Female	Asian	25	Biochemistry	Fluent (Mandarin); Beginning (Spanish)

Assessment-Focused Instruction

Through three program courses – *Learning Theories* (Summer 2010), *Science Education Theory* (Fall 2010), and *Science Methods* (Winter 2011) – the teachers participated in assessment-focused instruction led by me (the author). The instruction was (a) theoretically guided by the CUE framework, (b) designed to provide spaces for the teachers to reflect upon their assessment expertise, including beliefs about assessment, and (c) designed to support teacher learning and application of assessment in linguistically diverse science classrooms. Table 2 displays the specific course goals, assessment-focus goals, and sample activities/readings. By participating, the teachers had exposure to the aspects of science classroom assessment that I would analyze. The instruction also allowed me to observe expertise through course activity and products in a more ecologically valid environment, meaning that the teachers were observed and their products analyzed *as* they were learning about assessment.

Learning Theories was the first course the teachers took in the program, occurring in the summer just prior to the onset of their teaching practicum. Along with the other secondary preservice teachers, they were exposed to theories of schooling and learning, with a focus on teaching diverse students. Besides the weekly two-hour lecture, the teachers met in their cohorts for a weekly two-hour discussion, led by a teaching assistant, so that they could apply concepts learned during lecture to content-specific teaching. The assessment-focused instruction for this course consisted of two parts. First, in the capacity as a guest instructor, I led a two-hour

lecture during the fourth week, which included opportunities for the preservice teachers to discuss with other students and report out on their own experiences being assessed as students. The preservice teachers also read several assessment "vignettes" and engaged in a discussion to broaden their view about various assessment forms/functions and understand that at its core, assessment is a way to find out what students know and can do. Finally, I provided an overview of assessment principles and issues, related to each dimension of the CUE framework. Similar to other weeks, during the discussion section, a trio of preservice teachers led activities based upon the readings to delve further into course content. For their activity, the small group used a set of science assessment items to discuss the influence of language and culture on assessment, drawing on a reading by Solano-Flores & Nelson-Barber (2001). After the group led activity, I facilitated a discussion in which we further elaborated upon and clarified concepts from the lecture.

In the fall quarter (September to December 2010), the teachers took S*cience Education Theory* to examine theoretical approaches to the learning and teaching of science and to discuss ways in which to apply such theory into classroom practice. The assessment-focused instruction in this course aimed to expand upon the principles of classroom assessment in the context of *science* teaching and learning. Instead of devoting a single lesson to assessment, the instructor and I (as a volunteer graduate assistant) integrated assessment topics into class nearly every week; thus assessment became a course theme along with equity and scientific inquiry. For four of the weeks, the teachers participated in assessment case studies to learn about

various assessment forms (e.g., concept mapping, performance assessment) and to use their assessment knowledge to critique particular scenarios based on the Construction, Use, and Equity dimensions. For instance, the teachers watched a video clip of students participating in a science performance assessment to accompany a reading they had about the language demands of science performance assessments (Shaw, Bunch, & Geaney, 2010). Groups discussed the scenario using a list of prompts, which usually led to a whole class discussion. As homework, individual teachers then described what they would do differently in relation to what the assessment assessed, how it was interpreted, how information would be used to support learning, and how to make it more equitable for ELs.

In the winter quarter (January to March 2011), the teachers took *Science Methods* to focus on the *practice* of teaching science, drawing upon theory learned through other courses. In this course, I (again as a guest instructor) led two 90-minute assessment workshops. These workshops focused on applying what they had learned throughout the program to planning and using assessment in the context of their teaching practicum. In the first workshop, guided by a template, the teachers constructed (or modified) an assessment that they could use in their teaching practicum. The teachers brought in actual student work in the second work and, again using a template, reported on patterns they found from interpreting student work and described how they would give feedback and modify instruction. In both workshops, prompts on the template helped them consider equity issues for ELs.

Table 2

Assessment-Focused Instruction Objectives and Activities

Course objectives	Assessment-focused instruction objectives	Specific activities and selected readings
Learning Theories (Summer 2010)		
Engage students in the critical analysis of theory and practice in the teaching of subject matter to diverse secondary school student populations ***and*** apply the core theory to their specific subject teaching area	1. Understand a framework for conceptualizing educational assessment 2. Be exposed to the major issues in educational assessment 3. Begin connecting assessment to various learning theories	Lecture (with think/pair/share and discussion) Student-led activity (Black & Wiliam, 1998b; Shepard, 2000; Solano-Flores & Nelson-Barber, 2001)
Science Education Theory (Fall 2010)		
Examine theoretical approaches to the learning and teaching of science ***and*** use such theory to inform classroom practice	Understand assessment concepts in relation to science education (forms of science assessment, assessing scientific inquiry and practices, science standards, supporting science learning, assessing linguistically diverse students)	Mini-lectures Discussion Case studies Essays (equity reflection, research paper) Reading jigsaw (Furtak, 2009; Klassen, 2006; NCR, 2001; Shaw, Bunch, & Geaney, 2010; Siegel, 2007)
Science Methods (Winter 2011)		
Apply core theories of science education to the practice of science teaching	Apply assessment understandings to constructing an assessment of student learning and using assessment information to plan instruction	Two assessment workshops (model the process of constructing assessment/using assessment info and provide Teachers two templates Furtak (2009)

Method

Research Design

To determine the ways in which the teachers' assessment expertise changed when exposed to the theoretically guided assessment instruction, I employed a mixed methods triangulation design (see Figure 1). I collected and analyzed both qualitative (open-ended survey item responses, program products, and interview transcripts) and quantitative (Likert scale item survey responses, scored program products) data. In total, I developed and used two instruments and one scoring rubric. First, a set of teacher interview prompts solicited the teachers' assessment expertise through an interpretive epistemology, meaning that, guided by theory, I was open to patterns that emerged from the data rather than using an a priori coding scheme. Second, consisting of likert scale and two open-ended items, the Secondary Science Assessment Survey (SSAS) tracked measurable changes across time in the program. I transformed responses to open-ended items into quantitative data by scoring those responses with the CUE Assessment Expertise Rubric (see Chapter 1). I compared quantitative and qualitative changes in assessment expertise while interpreting the findings (thus mixing methods).

A pragmatic philosophical position guides the mixed methods approach in that the research questions ultimately drive the design and, if best suited for the design, multiple worldviews (as done in this study) might be called upon (Creswell & Plano Clark, 2007; Johnson & Onwuegbuzie, 2004). The use of a mixed methods approach has several advantages. Although most of the data were qualitative in nature, Weber

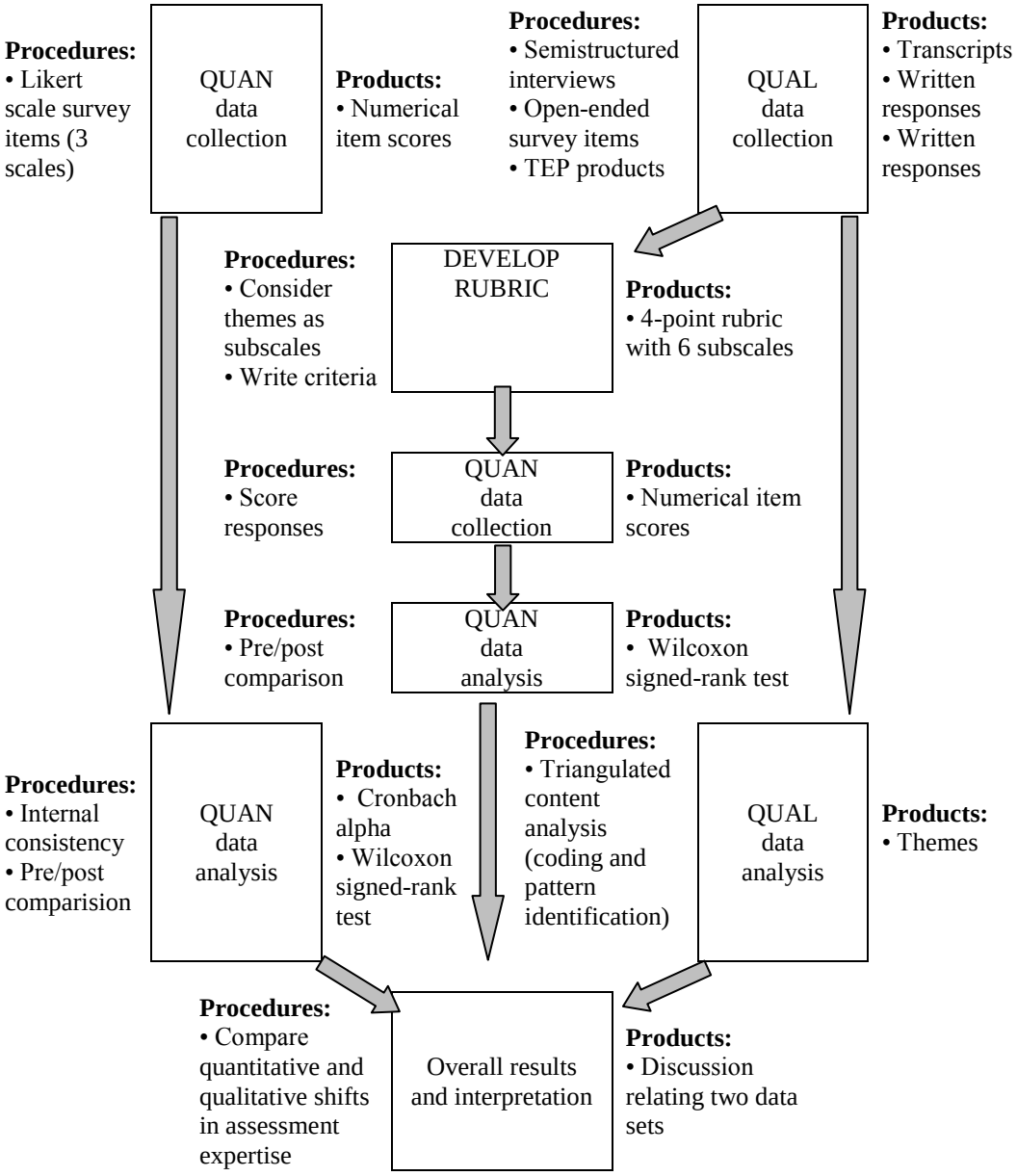

Figure 1. Sequential flow of the mixed methods triangulated research design (QUAN = Quantitative; QUAL = Qualitative). Boxed text indicates the various stages, such as data collection and data analysis. Text to the left of the boxes indicate procedures done during the respective stage (e.g., scoring responses), while text to the right of the boxes indicate the products (e.g., numerical item scores).

(1990) notes, “The best content-analytic studies use both qualitative and quantitative operations on texts” (p. 10). Furthermore, Morgan (2007) argues that a pragmatic approach moves beyond strictly inductive or deductive reasoning by operating under a version of abductive reasoning, in which the design iteratively cycles between induction and deduction and looks for the best possible explanation to the data presented. By triangulating different, yet complementary data, I brought together the strengths of both qualitative and quantitative analyses to enhance the validity of the results (Creswell & Plano Clark, 2007).

Data Sources and Measures

In a review of studies about teacher cognition, Kagan (1990) made four claims about appropriate measures to evaluate teacher beliefs, knowledge, and thinking. First, the most successful studies identify qualitative, as opposed to precise quantitative, descriptions. Second, it helps to “define desirable behaviors and cognitions in terms of a specific pedagogical framework” (p. 459). Third, multi-method approaches are superior, since they capture the complex aspects of teaching. Finally, it is difficult to establish ecological validity of techniques sensitive to short-term changes in teacher belief. Regarding the last point, Kagan defined ecological validity as “the kinds of evidence researchers provide concerning the relevance of a measurement technique to classroom life” (p. 422) and expressed particular concern whether “teachers’ performances on a particular tool or task related to their classroom behaviors or to valued student outcomes?” (p. 422). Despite the challenges in analyzing complex forms of cognition such as assessment expertise, I was able to

develop and use multiple instruments that would be the best suited for answering my research questions. Although I did not heed Kagan's caution about the limitations of quantitative measures, my collection of instruments were consistent with a multi (and mixed) methods approach that drew on a specific pedagogical framework (CUE framework) and included measures to enhance the ecological validity of the analyses (e.g., the Performance Assessment for California Teachers [PACT] commentary). Next, I describe the structure, purpose, and trustworthiness/validity of the data sources, outlined in Table 3.

Table 3

Data Sources

Onset of program (July 2010)	During program (August 2010 to April 2011)	Toward the end of program (May 2011)
Teacher interview 1	Teacher interview 2	Teacher interview 3
SSAS 1	SSAS 2	SSAS 3
	Learning Theories final project	PACT commentary

Secondary Science Assessment Survey. I developed the Secondary Science Assessment Survey (SSAS) as a mixed methods measure of changes in teachers' assessment expertise (see Appendix A). Drawing on items from other science teacher belief surveys (Genc, 2005; Stoddart, Bravo, Solís, Mosqueda, & Rodriguez, 2011), the SSAS consisted of 46 4-point (*strongly agree, agree, disagree, strongly disagree* or *very important, important, somewhat not important, not important*) likert scale items related to desirable science teaching practices. The first administration of the SSAS also included items that elicited demographic information, academic

background information, and experiences learning about educational assessment. An expert in assessment and ELs reviewed the item pool and changes were made accordingly. From the 46 items, I hypothesized that there would be three scales associated with the Construction, Use, and Equity dimensions of the conceptual framework. I also included items that related generally to equitable science teaching and teaching scientific discourse and argumentation, but these items were not analyzed and thus not reported in this study. In order to test the reliability of the three hypothesized scales, an additional 26 individuals – the entire secondary math cohort (N = 10) in the same program as well as 16 secondary teachers at two nearby universities (13 science, 3 math) – completed the likert scale survey items. I included math teachers given the small sample of available science teachers that I could recruit with limited resources. To account for differences in subject matter, I adapted the items administered to math teachers to relate to math teaching. Based upon this sample of 37, I confirmed the reliability of the three scales: Construction (7 items, α = .809); Use (6 items, α = .710), and Equity (9 items, α = .740). Scales with a Cronbach's alpha of above .700 have acceptable reliability, while those above .800 have good reliability (George & Mallery, 2003). Although not done in this study, a confirmatory factor analysis will be conducted in continuing research to validate the scales with a larger pool of teachers.

The SSAS also included two open-ended items that elicited how the teachers applied assessment understandings while engaging in two distinct tasks. The first open-ended item, focusing on *assessment planning*, asked the teachers to (1) choose

one of the following science topics – Mendelian genetics, acids and bases, light and optics, or earthquakes, (2) "describe in as much detail as possible how you would assess student learning during this unit," and then (3) "explain why you would assess this way." The second open-ended item, focusing on *assessment critiquing*, presented a hypothetical vignette about how "Ms. Sanchez" assessed her students during a particular set of lessons (see Appendix B). The prompt asked the teachers to (1) describe and explain to what extent they thought Ms. Sanchez's assessment practices were effective, including specific things they thought were or were not effective and (2) describe what they would do differently, and then (3) list other information (if any) they would like to have about the scenario to comment on her assessment practices. During the first interview (see teacher interviews below), I asked each teacher to go through their critique to check whether they understood the scenario and directions and to check whether the interview probing yielded additional information not captured by the written response. Responses during the interview suggest that each teacher did understand the scenario and that the written response sufficiently allowed them to convey their critique. Moreover, I asked a similar question to the assessment plan prompt during interviews as a form of triangulation.

Teacher interview protocols. As stated by Patton (2002), "qualitative interviewing begins with the assumption that the perspective of others is meaningful, knowable, and able to be made explicit" (p. 341). While the surveys allowed me to report general, measurable patterns of assessment expertise, interviews allowed me to use an interpretive worldview to, in their own words, understand the teachers'

assessment expertise. Drawing upon previously used interview protocols eliciting assessment beliefs and practices (Lyon, 2011), I developed three semi-structured teacher interview protocols (see Appendix C). The first and third interview protocols included similar prompts to compare assessment expertise at the beginning and end of their program in relation to the three dimensions of the CUE framework. Some prompts solicited assessment understandings (e.g., "what does it mean to you to equitably assess student learning?), whereas other prompts asked participants to hypothetically articulate how they would assess student learning and use such assessment information.

Teacher products. I also analyzed and scored two teacher products, the teachers' final project for the *Learning Theories* course and their written commentary to the Performance Assessment for California Teachers (PACT). The two teacher products, being embedded with the program curriculum and required for course and program completion, served as more ecologically valid measures at two different time points (*Learning Theories* final project – August of 2010; PACT – May 2011). In the *Learning Theories* final project, the teachers wrote a description of three activities that would teach a particular science standard and as well as two assessments of the learning objectives affiliated with the standard. The teachers were expected to describe how the activities drew upon theories of learning, to identify language demands of the activities (including assessments), and to describe how the activities were responsive for diverse learners.

The PACT[4], is a culminating teacher product used by thirty universities across California to determine whether teachers have demonstrated proficient competencies associated with the California teaching standards. To complete the PACT, teachers first plan and implement approximately a week's worth of lessons, video tape two self-chosen segments of these lessons, and write an extensive commentary that addresses five teaching areas – Classroom Context, Planning, Instruction, Assessing, and Reflecting. A set of prompts and rubrics guided each teaching area. To examine a comparable task to the *Learning Theories* final project and make scoring more reliable, only the responses to the "planning" prompt were analyzed. Moreover, while responding to the planning prompts, the teachers were instructed to articulate how their instruction and assessment addressed the needs of diverse learners, with a particular focus on academic language.

CUE Assessment Expertise Rubric. To measure assessment facility in the Construction, Use, and Equity dimensions, the 4-point CUE Assessment Expertise Rubric was developed and used to score (a) the three SSAS assessment plan responses, (b) the *Learning Theories* final projects, and (c) the PACT planning commentaries. Higher points on the rubric corresponded to higher levels of expertise. To increase the clarity of constructs and measurement reliability, each original scale (Construction, Use, Equity) was divided into two sub-scales (Construction – *assessment task* and *alignment*; Use – *curricular context* and *cycle of inquiry*; Equity – *fairness* and *access*) (see Chapter 1 for details).

Data Collection and Analysis

Data collection. As indicated in Table 3, the SSAS was administered to the teachers at three time points during the program – at the onset (July 2010), middle (December 2010), and toward the end (May 2011). In the weeks following each administered survey, I interviewed each teacher. I collected an electronic copy of each teacher's *Learning Theories* final project at the end of their *Learning Theories* course (August 2010) and collected the PACT commentary at the time it was due to the university (May 2011). Throughout the year, I collected additional course products, which I did not analyze systematically for this study, but did inform my thinking when analyzing the other data sources.

Qualitative analysis. To analyze qualitative data, I primarily engaged in content analysis, defined broadly as ''any qualitative data reduction and sensemaking effort that takes a volume of qualitative material and attempts to identify core consistencies and meanings'' (Patton, 2002, p. 453). I conducted the content analysis of written material through iterative rounds, both during and after data collection, of coding and pattern recognition to formulate "patterns" that qualitatively capture changes in assessment expertise (Saldana, 2009). I first engaged in a round of structural coding, in which I organized each data source into textual excerpts that aligned with one of the three framework dimensions. I uploaded the textual excerpts, structurally coded, into a hyperRESEARCH database (http://www.researchware.com/hr/index.html; Hesse-Biber, Dupuis, & Kinder, 1991). I then engaged in several rounds of more descriptive coding of the excerpts, resulting

in a coding scheme (see Table 4). The main purpose of the first round coding was to identify various elements of assessment expertise, grounded by the CUE framework, which could indicate understanding of and facility with assessment.

Table 4

Excerpt from Coding Scheme

Code	Description	Example
	Construction	
Assessment task	Tasks identified as an assessment or used by the teacher to find out what students know and can do	"All these different types of assessment like portfolios, and notebooks and tests and essays and performance based"
What is being assessed	The content of assessment or the type of knowledge/ skills/ attitudes/etc. being assessed	"Practice quizzes with that sort of difficult logical thinking and then have students work together to solve those"
Science learning	How students learn science or theories of learning that guide assessment construction	"I feel like the biggest one is hands on learning"
Alignment	Alignment between assessment task and (a) learning objectives, (b) evaluative criteria/method of interpretation, and/or (c) learning theory	"I really think that depending on what kind of content I am teaching and the way that I am presenting the information is the way that I would construct the test"
Criteria	References to rubrics or other means to mean to interpret student work	"What I have learned is that you got to be fair in evaluating students and its easier to have a rubric when you are grading or you are comparing"
	Use	
Assessment purpose	The role of assessment at the classroom level	"Assessing a student is you are seeing what they know and I guess

		what they don't know, like how much they're grasping the concepts of the class"
Assessment placement	When or how often to assess during the curriculum	"So maybe like assessment at the beginning would be as students explore and play with lenses and lasers and things"
Formative assessment strategies	Ways in which to provide feedback, modify instruction based on assessment information, or have students assessment themselves/peers	"Well I usually provide feedback right on their assessment so I let them know what it is they're missing or what it is that they need"
	Equity	
Equity perspective	Explicit references to what equity means or what it means to equitably teach/assess	"Equity teaching means that I can bring everyone to the same level no matter what kind of background they have"
Assessment fairness	Ways in which to ensure all students, regardless of language proficiency, have an opportunity to fully demonstrate what they have learned	"I am trying to be as visual as possible so you don't really even need language to understand the problem"
Access	Ways in which students can access the science curriculum by promoting full participation, use of language, or complex thinking	"And so, that's kinda been my goal. Is to… uh, try to focus on like arguments um… and other… mostly arguments but like other kinds of academic language"

Using the coding scheme, I reanalyzed data by finding excerpts from each data source of each participant to exemplify each code. For the three time points, I reorganized and reanalyzed coded data to develop longitudinal patterns related to each dimension of assessment expertise (Saldana, 2009) using the constant-comparative method (Corbin & Strauss, 1990). Throughout the entire process of data

collection and analysis, I wrote research memos to synthesize procedures and coded data and aid in presenting a transparent path of analysis from raw data to inferences drawn (Corbin & Strauss, 1990).

Quantitative analysis. I quantitatively analyzed (a) likert scale item responses, (b) open-ended responses, (c) content from the *Learning Theories* final project, and (d) content from the PACT planning commentary.

For hypothesized likert item scales, I checked their internal consistency by finding the Cronbach alpha statistic; dropping items as needed to ensure the statistic was above .700 (see instruments and measures section). I computed each of the three scaled scores (Construction, Use, Equity) by averaging each teachers's responses (1-4) to items in the corresponding scale.

To score the qualitative products with the CUE Assessment Expertise Rubric, I trained two graduate students with classroom experience and researcher interests related to science/teacher education. By collectively and independently scoring practice responses, we reached consensus about particular interpretations of the rubric criteria and changed the rubric as needed to aid in scoring. I blinded each response by participant and time point, and I randomly assigned each response to two of the three scorers. Thus, the scorer knew the data source (e.g., survey assessment plan, PACT commentary), but not who wrote the response, or in the case of the assessment plan and critique, when it was written. Initial scoring yielded the following average percent agreement and Pearson's correlation values (respectively) for each scoring pair across all six sub-scales: *Scorer1-Scorer2* (53%, .469); *Scorer1-Scorer3* (52%,

.344); *Scorer2-Scorer3* (49%, .502). Although the agreement is moderate at best, due the relatively small number of items scored (68, not counting the 15 used as anchor responses), all scorers were able to meet again to discuss and reach consensus on all responses with a 2-point or more discrepancy between scorers, which also helped refine the scoring criteria and decisions. I then took the role of an expert scorer and, based upon refined decisions, independently scored responses with a 1-point discrepancy.

For all quantitative data, I examined descriptive statistics to uncover general patterns in distribution, cohort mean and spread at each time point and over time. Given the small sample size, the intent of the study was not to generalize to a population larger than the sample. Therefore, I used a Wilcoxon signed-rank test (non-parametric analog to a paired-sample *t* test) to test for differences in the central tendency of means between the start and end of the program (Shavelson, 1996). Quantitative results in the form of score distributions and frequencies were compared with qualitative patterns as a way to triangulate findings and provide a more complete picture of the assessment expertise changes. I next report the findings from analyzing the data.

Broad Patterns of Change in Assessment Expertise

In this section, I report on the broad patterns of assessment expertise change, as measured longitudinally by the SSAS. Teacher scores from the *Learning Theories* final project and PACT planning commentary complement the broad analysis. The patterns reported will then be explored further in subsequent sections.

Assessment Understanding

Figure 2 displays the teachers' assessment understanding over the span of the program as measured by the SSAS likert item scales. The scales reflect, on average, to what extent teachers agreed with the assessment tenets and practices outlined in the CUE framework. Descriptively over time, the average scaled score along each conceptual dimension increased; however, the patterns of change varied among the three dimensions. At the beginning of the program, the teachers demonstrated greater assessment understanding related to the Construction dimension (M = 3.38 out of 4, SD = .266) than with the Use dimension (M = 3.26, SD = .216) and the Equity dimension (M = 3.22, SD = .243). However, by the second administration of the survey (four months later), the variation in means dissipated (Construction: M = 3.54, SD = .337; Use: M = 3.59, SD = .313; Equity: M = 3.56, SD = .258), a trend that continued into the last administration, another five months later. A Wilcoxon signed-rank test revealed that the change between the first and last administration was statistically significant for the Use dimension (p = .032) and the Equity dimension (p = .014), but not statistically significant for the Construction dimension (p = .082).

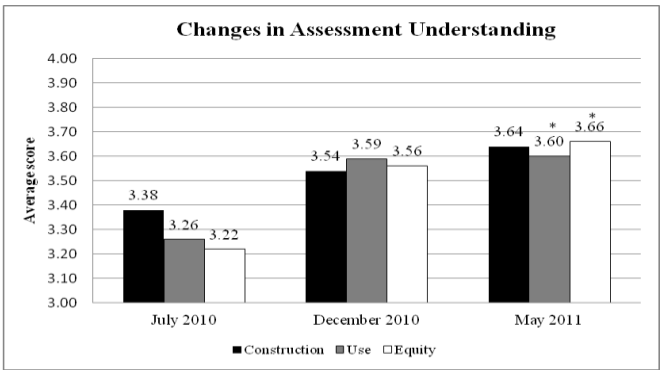

Figure 2. Average change in the teachers' assessment understanding scores (based on three 4-point likert scales) across the three time points (July, 2010; December, 2010; May 2011). * denotes statistical significance at $\alpha = .05$.

Assessment Facility: Critique and Planning

Two data sources measured assessment facility over time in the program. First, Figure 3 displays the average assessment plan scores, in which the teachers articulated how they would assess learning during a selected science unit. Similar to the assessment understanding pattern, the teachers demonstrated positive changes throughout the program. Moreover, during the first administration, the teachers scored the highest on the Construction dimension ($M = 4.09$ out of 8, $SD = .831$), although they scored higher on the Equity dimension ($M = 3.09$, $SD = 1.14$) than on the Use dimension ($M = 2.91$, $SD = 1.22$) unlike the initial assessment understanding pattern. Again descriptively, the assessment plan scores continuously increased over the span of the program, whereas the assessment understanding scores did not increase as much from the second to third administration. However, while the teachers changed *the most* in their assessment understanding related to the Equity dimension, their application of this understanding changed *the least* in the assessment plan. A Wilcoxon signed-rank test revealed that the change between the first and last administration was statistically significant for the Use dimension ($p = .005$), but not statistically significant in the Construction dimension ($p = .058$) or the Equity dimension ($p = .130$).

Figure 4 displays the average assessment critique scores, in which the teachers critiqued a hypothetical assessment scenario (see Appendix B). The patterns were similar in some aspects to the assessment plan scores. In the first administration, teachers scored highest in the Equity dimension and the lowest in the Use dimension. Furthermore, a Wilcoxon test revealed that statistically significant changes occurred only in the Use dimension (p = .036) over time, similar to changes in the assessment plan scores. Like the assessment plan scores, the teachers' assessment critique scores increased over time in the Use and Equity dimensions.

Yet, some patterns were different. Not only were changes in the Construction and Equity dimensions statistically significant (p = .429 and p = .070, respectively), the average Construction score was actually lower in the second and third administration (3.30 and 4.00, respectively) as compared to the first administration (4.27). Finally, both Use and Equity dimension scores were considerably higher while critiquing (3.82 and 4.36, respectively) than while planning (2.91 and 3.09, respectively).

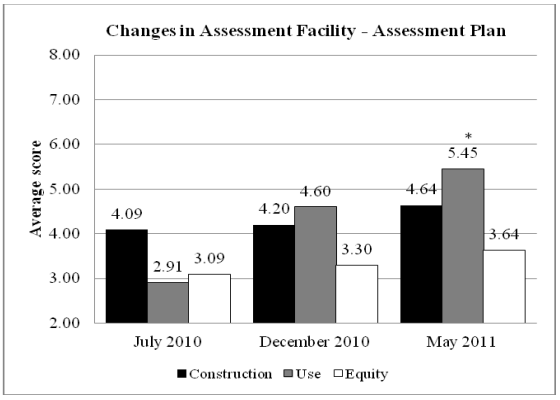

Figure 3. Average change in the teachers' assessment facility scores (assessment plan) across the three time points (July, 2010; December, 2010; May 2011). * denotes statistical significance at α = .05.

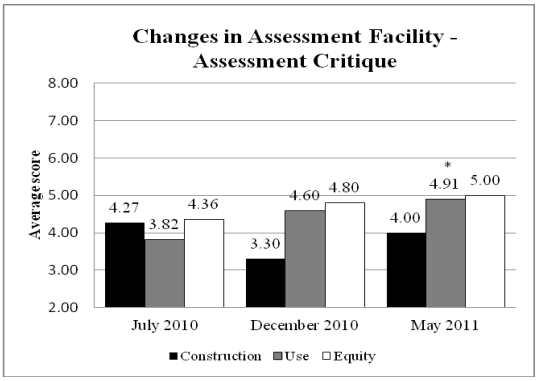

Figure 4. Average change in the teachers' assessment expertise scores (assessment critique) across the three time points (July, 2010; December, 2010; May 2011). * denotes statistical significance at α = .05.

Scoring the *Learning Theories* final project and the PACT planning commentary provided an alternative measure of assessment facility, one in a more ecologically valid context. Due to differences in the prompt, evaluative criteria, and context in which teachers complete each task, temporal comparisons cannot be made between the two sources; yet it is important to note that the scores represent assessment facility at two time points after the start of the program (*Learning Theories* final project – six weeks; PACT planning commentary – 10 months).

In comparison to the SSAS assessment plan scores, both situated assessment plan scores were, across all dimensions, higher. While in the first assessment plan the Construction score was considerably higher than the Equity score, in the *Learning Theories* final project (again completed between first and second administration of

the SSAS), the Equity score was actually higher. However, the pattern of PACT planning commentary scores across dimensions (completed just prior to the last administration of the SSAS) was consistent with the third SSAS assessment plan scores – highest in the Use dimension and lowest in the Equity dimension.

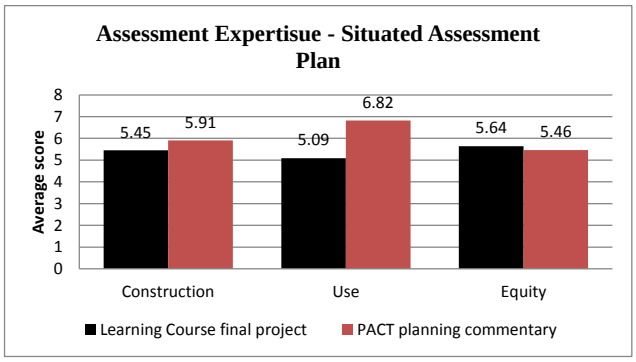

Figure 5. Average change in the teachers' assessment facility scores based on responses to two assessment plans situated within the program requirements – their *Learning Theories* final project and their PACT planning commentary.

To summarize, the descriptive statistics reported indicate that, on average, the teachers do change in both assessment understanding and facility over time in the program, although the Wilcoxon test revealed that changes between first and third administration were only statistically significant in the Use dimension. More importantly, noticeable patterns emerged when looking across the three dimensions – warranting further exploration about changes through quantitative and qualitative analyses. A closer look at the distribution of scores and analysis of written content (e.g., interview transcripts), reported next, can provide insight into change not captured through simple changes in means. Only changes in the assessment plans will

be reported systematically to represent assessment facility, although changes in the assessment critiques will be referred to from time to time as supporting evidence.

Construction Dimension: Expanded Assessment Task Repertoire *and* Aligning Assessment Task to Learning Objective

The Construction dimension explored the teachers' expertise in constructing (or selecting) assessment tasks that elicit student thinking and that are theoretically cohesive (as laid out in the assessment triangle model). Two overarching patterns emerged during the teacher interviews: the teachers *expanded their assessment task repertoire* and *considered the alignment of assessment tasks with learning objectives*. Both patterns were reflected in the teachers' assessment facility.

Assessment Understanding

Over the span of the program, the teachers explicated knowledge of a more expansive repertoire of tasks that they could use to assess science. As Darlene expressed, "I have a lot more variety, ways of assessment in my bag, and I can use all these different ways that I haven't even thought about before" (Interview 2). For example, when asked about knowledge of different assessment tasks during interview one, Darlene responded with "exams, standards, tests, quizzes," whereas during the third interview she listed the following assessments:

> Warm up problems, in class discussion, warm up with groups, labs are the big assessment, homework, test obviously, you don't see projects that much but it is definitely, you can do it and if you have a lot of time, you can have them teach and have inquiry labs, in class discussion like, "how many of you think that this is exothermic or this is endothermic?"

In the example, Darlene does not articulate *how* the specific tasks will be used to assess science. Yet, by recognizing that assessment can take various forms, she is better equipped with flexibility while selecting tasks best suited for the content she wants to assess.

At the beginning of the program, despite limited knowledge of assessment tasks, the teachers understood that science assessment should elicit students' scientific thinking (e.g., problem solving, explaining scientific phenomena), not just factual recall. However, by the end of the program, many teachers noted that assessment can also capture students' *use of language* in science, something they did not refer to in the first interview. For example, Whitney described how the assessments she employed during her teaching practicum would "focus on getting them [students] to use that scientific language in their explanations" (Interview 3), while Dean articulated how he would "try to focus on…mostly arguments but like other kinds of academic language" (Interview 3). Thus, the teachers' descriptions of *what* to assess and what assessment *forms* to use changed. Another aspect of the Construction dimension, though, is to link the content and form of assessment (i.e., the task) to underlining theory about student learning.

The assessment triangle (NCR, 2001) was employed as a model to analyze the teachers' expertise in linking three components of assessment design – cognition, observation, and interpretation. In relation to these components, the teachers shifted from understanding assessment as an isolated task to considering the alignment between the task and the instructional learning objectives. As Lauren described in the

second interview, “I definitely feel like I’m way more informed to go through that [assessment] process. Like once I have a learning goal, I feel like I know I have a better place of where to start to think about the assessment.” Her understandings about assessment in relation to an objective carried over while reflecting on her own teaching practices, also representative of the teachers:

> Ideally...I would start with the assessment...I would ask myself okay so what do I want them to remember in ten years, and what do I want them to be able to do and start there and design the assessment that's going to get at that and then use that to help me plan my learning activities and how I would teach the content. (Interview 3)

While describing the backward design model (Wiggins & McTighe, 2005), Lauren did not mention or allude to assessment for evaluative purposes. Instead, she considered the coherence between task and learning objectives (or goals) because that would better allow assessment to inform her instruction and, thus, aid in student learning.

According to assessment triangle model, not only should the assessment tasks align with how student learn and what is important to learn, but the tasks should also align with the evaluative criteria used to interpret student work. Teachers were exposed to rubrics as a way to construct and organize evaluative criteria throughout their program, including the assessment-focused instruction. While the teachers, generally, had heard about rubrics before the program, as stated by Darlene in the second interview, “Once I came [in]to the program, everything was off of a rubric.”

The teachers began articulating the roles and limitations of rubrics. For example, during the second and final interview, nearly all teachers expressed a view

about rubrics, primarily positive, or described how they have or would have used them:

> For me it [a rubric] makes it really clear when I'm looking through student work like okay this is what I expect and really breaking it down like, okay… what is each thing [criterion] worth and…it's not like not judgment, but no bias I guess. It's like they did not meet this even if I think the student didn't do as well as I think they could have or like whatever reason. It's like this is where they fell in the rubric and it's not like bias on my part. (Glenda, Interview 3)

However, teachers were less likely to consider the alignment between criteria in the rubric and the learning objectives explicitly. Instead, they demonstrated a more general use of rubrics, often due to their perceptions about a rubric's limitations. As noted by Willow, "Content wise, the rubric wouldn't really help at all but if I wanted structure then the rubric would help…My rubrics are not about content" (Interview 3). By structure, Willow was referring to the organization of texts, such as the structure of a lab report conclusion (did it have a claim, relation to hypothesis, evidence, etc.). When asked if she would use a rubric to interpret what students know and can do, Teresa, responded, "No I would not because the rubric…it's more concrete in that you're looking for certain things like did you include this" (Interview 2). Although they understood how rubrics can be helpful, the teachers were hesitant to include criteria that focused on more than just the presence of particular vocabulary or structural elements, often resulting in a rubric that appeared to lack alignment with the learning objectives in practice.

Assessment Facility

Figure 6 represents changes in each teacher's assessment plan score, across all three dimensions. Each score has two components (the sub-scales), which is why the graphs are two-dimensional. Larger "bubbles" represent more teachers with that coordinated score. Thus, the figures collectively represent changes in score distributions. As indicated in Figure 6, by the third assessment plan, all teachers either shifted toward higher levels on the assessment task sub-scale or remained stable at level 3, representing assessments tasks that elicited scientific thinking in some capacity. Besides Darlene, who fluctuated through the three plans and Hallie who remained stable, all teachers reached level 2 expertise in the alignment sub-scale. This patterns indicates that they considered an alignment between assessment task and learning objective. Ten teachers included some recognizable assessment task on the first assessment plan, but on the second and third assessment plans, they described a greater *variety* of assessment tasks and described those tasks in more depth (two elements *not* captured by the rubric). By the third assessment plan, 9 of the 11 teachers demonstrated level two expertise in both sub-dimensions – only three of which started off that way.

Darlene demonstrated the most growth in her *assessment task* expertise level, from level one (assessment plan 1) to level two (assessment plan 2) to level three (assessment plan 3). On the first assessment plan, she described what she would assess, related to acids and bases, but on the third plan, she listed (although not described in depth) *eight* distinguishable assessment tasks (e.g., practice problems, concept map, lab).

The shift from isolated to connected assessment task was evident while comparing Laurens's first and third assessment plan. In the first plan, she did not reference the learning objectives (i.e., what she wants students to *learn*):

> I would give them [students] several sets of parent alleles (ie. 3 pairs of parents) and have them make a punnet square for each parent set and have them determine the phenotypic ratios and have them figure out what combinations of parent alleles would produce that ratio outcome.

However, during the third assessment plan, Lauren not only described the assessment task, but also connected the task to the learning objectives: "… This [assessment] would also tell me if students understand the concept of dominant & recessive alleles and the difference between a genotype & phenotype."

What appeared absent in the assessment plans is a connection between assessment task and theoretical underpinnings of those tasks. In other words, when describing assessment tasks in interviews or assessment plans, they rarely mentioned the theory of learning (e.g., behaviorist? constructivist? socio-cultural?) or explicate the importance of assessing particular science content. During assessment-focused instruction, students read about the theoretical underpinning of assessment forms, and we discussed the importance of such underpinnings (see Furtak, 2009; Klassen, 2006; NCR, 2001; Shaw, Bunch, & Geaney, 2010).The rubric for the *Learning Theories* final project and the PACT commentary explicitly called upon the teachers to relate learning/assessment tasks to theories of learning. While some teachers made connections in those two tasks between assessment and theoretical underpinnings, they emphasized the connection between theory and instructional practices more so than assessment tasks.

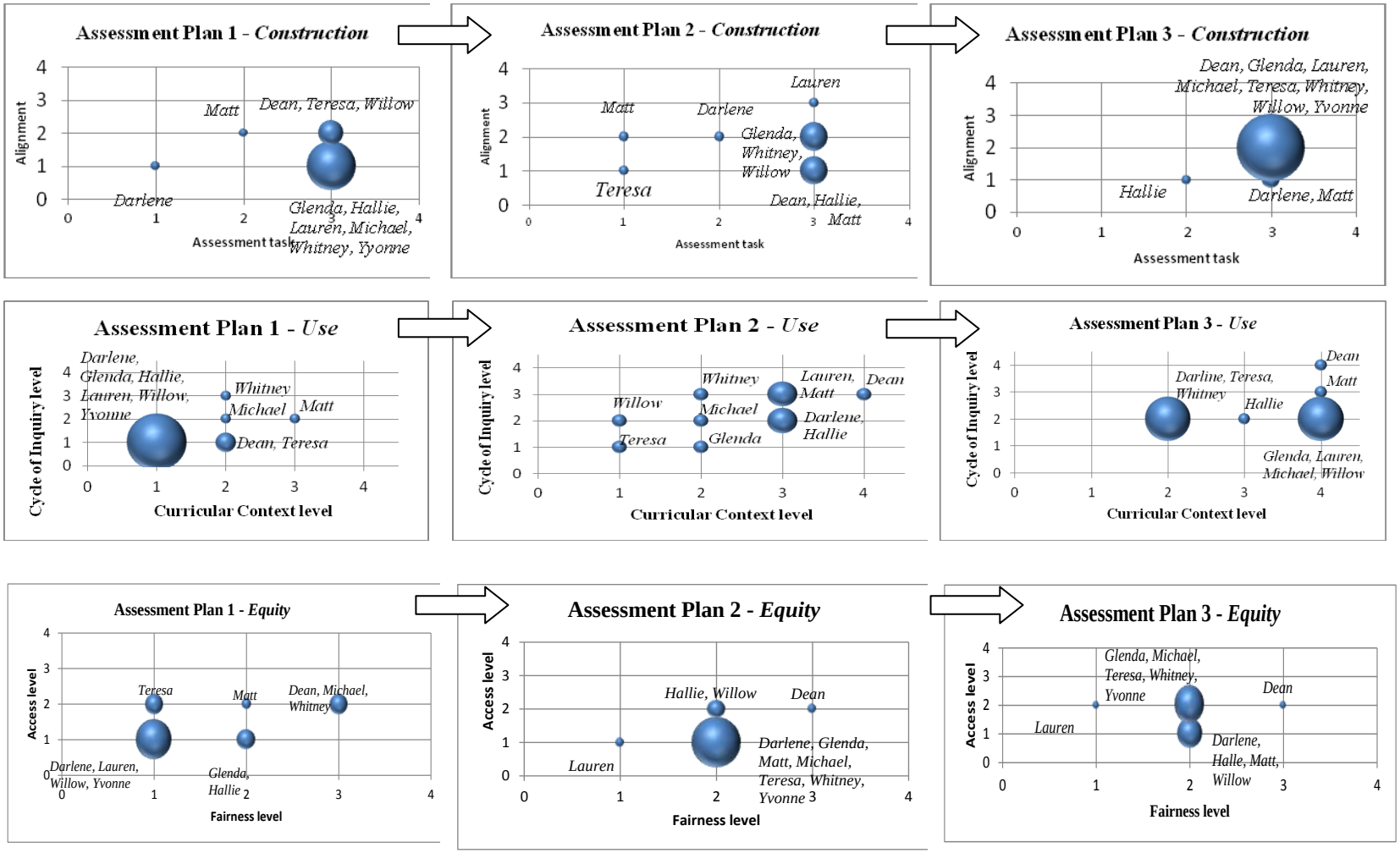
Assessment Plan 1 - Construction
Alignment
Assessment task
Matt
Dean, Teresa, Willow
Darlene
Glenda, Hallie, Lauren, Michael, Whitney, Yvonne
Assessment Plan 2 - Construction
Alignment
Assessment task
Lauren
Matt
Darlene
Glenda, Whitney, Willow
Teresa
Dean, Hallie, Matt
Assessment Plan 3 - Construction
Alignment
Assessment task
Dean, Glenda, Lauren, Michael, Teresa, Whitney, Willow, Yvonne
Hallie
Darlene, Matt
Assessment Plan 1 - Use
Cycle of Inquiry level
Curricular Context level
Darlene, Glenda, Hallie, Lauren, Willow, Yvonne
Whitney
Michael
Matt
Dean, Teresa
Assessment Plan 2 - Use
Cycle of Inquiry level
Curricular Context level
Whitney
Lauren, Matt
Dean
Willow
Michael
Darlene, Hallie
Teresa
Glenda
Assessment Plan 3 - Use
Cycle of Inquiry level
Curricular Context level
Dean
Darline, Teresa, Whitney
Matt
Hallie
Glenda, Lauren, Michael, Willow
Assessment Plan 1 - Equity
Access level
Fairness level
Teresa
Matt
Dean, Michael, Whitney
Darlene, Lauren, Willow, Yvonne
Glenda, Hallie
Assessment Plan 2 - Equity
Access level
Fairness level
Hallie, Willow
Dean
Lauren
Darlene, Glenda, Matt, Michael, Teresa, Whitney, Yvonne
Assessment Plan 3 - Equity
Access level
Fairness level
Glenda, Michael, Teresa, Whitney, Yvonne
Dean
Lauren
Darlene, Halle, Matt, Willow

Figure 6. The nine graphs indicate changes the distribution of assessment plan scores across time in the program. Each row of graphics represent a different conceptual dimension (Construction, Use, and Equity, respectively). For each graph, the score is represented as a coordinate of two sub-scores. The relative size of the "bubble" indicates the number of teachers with that particular coordinated score. For instance, in the top left graph, Darlene received a 1 in the "assessment task" sub-score and a 1 in the "alignment" sub-score, which is why there is a small "bubble" at the coordinate (1,1); whereas six teachers scored 1 in the "assessment task" sub-score and a 3 in the "alignment" sub-score, which is why there is a larger "bubble" at the coordinate (1,3). To examine changes over time, look at the graphs in each row from left to right.

Finally, only two teachers even mentioned rubrics on the assessment plans, both times during the second plan. Yet, the two teachers discussed rubrics in the context of setting expectations for students, not as a way to interpret what students know. However, seven of the teachers referred to rubrics while responding to the survey assessment scenario by critiquing Ms. Sanchez's decision to only use rubrics to score responses and not let students see the rubric (before or afterward), and by indicating that they would like to know specifics about the criteria in the rubric. Thus, they explicated some knowledge about rubrics at the onset of the program, but this knowledge did not translate to their assessment plans.

In summary, the Construction dimensions explored the teachers' expertise in constructing (or selecting) assessment tasks that elicit student thinking and are theoretically cohesive. The teachers expanded their assessment task repertoire and

considered the alignment between assessment tasks and learning objectives, although they did not fully consider the alignment of evaluative criteria to the task and learning objectives nor reference theoretical underpinnings of how students learn.

Use Dimension: Shifting toward a Formative Use of Assessment

The Use dimension explored the teachers' expertise at using assessment to support students' science learning; in particular, (a) when and for what purpose assessment was used within instruction and (b) the specific actions, such as feedback and modifications to instruction, considered by the teachers to help students achieve the planned learning objectives. Overall, the teachers began viewing assessment as a way to inform teaching and learning, not just evaluating it, and became increasingly specific about actions to support science learning – two patterns consistent with a shift toward formative use of assessment.

Assessment Understanding

As mentioned earlier, teachers' views of assessment, including its purpose, are central in preparing assessment-capable teachers (Abell & Siegel, 2011; Black & Wiliam, 1998b; Czerniak & Lumpe, 1996; Hill, Cowie, Gilmore, & Smith, 2010). At the onset of the program, the teachers collectively held a range of views about the purpose of assessment, the most common being to evaluate learning: "[The purpose of assessment is for] testing or making sure students are understanding…assessing how you're teaching…how you're getting the information across" (Hallie, Interview 1). Also expressed by other teachers, Hallie viewed assessment as way to evaluate herself, not just her students. Although less prominent, some teachers did

simultaneously view assessment's purpose as tracking student progress or informing teaching, consistent with a formative use of assessment.

Throughout the program, the teachers' evaluative view of assessment lessened as they became more inclined to view assessment as uncovering student learning/progress and using assessment information to support learning. In fact, all teachers professed during the second interview that they shifted toward a formative view of assessment:

> In the beginning, I was like, assessment is [a] test right? And you just give them to your students, and it's how they get their grades, and I wasn't thinking at all of that, you would use this as feedback on your...like to really look back at yourself and, um, reflect on what worked and what didn't, and what you need to go over more with your students. (Hallie, Interview 2)

Assessment became more than just an end of the unit phenomenon, which so many teachers alluded to when describing their experience as science students. Over time, the teachers recognized that assessment should occur continuously throughout the unit and include informal assessment that guides daily instruction:

> I would start by saying assessment isn't just the end of the unit test that you take... Assessment is happening all day every day…roaming around the classroom, checking on student progress, that's assessing. Reading their journal prompts, that's assessing. (Lauren, Interview 2)

Although less prominent, some teachers also explicated knowledge that assessments should ideally build on each other throughout instruction leading to the overarching learning objective – a concept explicitly discussed as part of the assessment-focused instruction during the *Science Method* course.

During the first interview, the teachers generally expressed positive attitudes toward providing students with feedback, modifying instruction based on assessment

information, and engaging students in self-assessment. The teachers talked generally about how they could provide feedback by talking to students, going over tests individually, or as a class, or writing comments on the assessment task. The teachers also tended to express that they *would* modify instruction, but as stated by Teresa, "bottom-line is I would do something different in class, I just don't know what" (Interview 1). Finally, while the teachers uniformly agreed that self- and peer-assessment are *good* practices, they professed being hesitant about incorporating such practices, particularly around concerns of fair grading. When probed, teachers often indicated that if they were not going to assign an actual score or grade then they would be more inclined to engage students in self- or peer-assessment.

Over the span of the program, the teachers changed in that they described the aforementioned actions in more depth and often discussed them in the context of their student teaching experience:

> The immediate feedback from the texting quizzes was good. Comments on their lab reports... I haven't used a lot of rubrics this year but I intent to and I really like the idea of having you know like filling out a rubric. Well for like major assignments that are graded where you fill out the rubric and give the rubric and the assignment back to the student so that they can... You know check their work against their rubric. I like to... It'd be neat to let them self assess themselves using rubrics too. (Lauren, Interview 3)

In this example, Lauren described feedback beyond written comments and individual discussions with students. She specified how to give feedback via "texting," a technology similar to clickers and described how feedback, including self-regulatory feedback can be enhanced through rubrics.

Assessment Facility

The changes in assessment understanding described above translated into similar changes in the teachers' assessment facility. Referring back to Figure 6, on the first assessment plan, a majority of the teachers mentioned an assessment task without describing explicitly *when* the task would occur or the *purpose* it would serve in the curricular unit, evident by 6 of the 11 teachers scoring a one on both Use sub-scales. From the first to third assessment plan, nine teachers increased in their curricular context score, with six teachers demonstrating level four expertise. Furthermore, the distribution of assessment plan shifted to a range of scores on the second assessment plan to two main clusters on the third assessment plan. While both clusters represented, generally, that the teachers incorporated assessment strategies that could support science learning, only one cluster of teachers (Glenda, Lauren, Matt, Willow) articulated the purpose and placement of multiple assessments instead of making a general reference to assessment as part of the curricular unit (Darlene, Teresa, Whitney, Yvonne).

On the first assessment plan, only one teacher described a specific strategy for providing feedback (discussing results from an in class activity related to the pH scale), while no teachers discussed how they could modify instruction based upon assessment or provide opportunities for self-assessment, a central feature for using assessment formatively. However, on the first assessment critique, seven teachers commented on how Ms. Sanchez should have provided feedback or use assessment to inform her teaching. Similar to teacher expertise with rubrics, while the teachers

explicated some recognition of formative assessment practices, they did not fully demonstrate expertise in applying knowledge in assessment plans.

Dean represented a teacher whose assessment plan changed considerably over the span of the program. While he scored a 3 out of 8 on the first assessment plan, just outside of the main cluster of teachers who scored a two, he scored an 8 out of 8 on the third assessment plan, the only teacher to do so. Table 5 compares Dean's first and third assessment plan.

Table 5

Comparison between Dean's SSAS 1 and 3 Assessment Plan Response

SSAS 1 Assessment plan response	SSAS 3 Assessment plan response
• Ask questions about the ways particles & waves move. Try to see if students can lead a reasoning of constructive & deconstructive interference of water waves → if light's a wave, what happens when two waves meet? • After an explanation of light in different mediums (snells law/reflection/refraction), multiple choice illustrations of reflecting or refracting light pathways • Multiple-choice/chose the correct image illustration correct pathway of light through a lense • Test students ability to use induction, apply-general-rules-to-specific-scenario type question. I think this would be an ending assessment, higher level.	In a unit on convex & concave lenses, I would assess S's learning in multiple ways. I would start with a lab where Ss find the image of an object refracted through convex & concave lenses. I would work with students in small groups and ask them to explain their observations, as well as questions such as: is the image inverted or upright; is the image larger or smaller than the object, is it real or upright; could you summarize what happens to the image as the object gets closer to the lens. These Q's are meant to discover commonalities in S thinking to guide formative assessment, & to lead Ss to consider key ideas. I would follow up the lab with a think-pair-share based on the previous questions. The "share" is an informal yet formative assessment in that I get to hear how Ss think about the key ideas, and at the same time scaffold content and language through my responses. At this point, a quick quiz where Ss respond by

	completing a diagram could be administered to check for understanding in a more visual and less language-dependent fashion.

In Dean's first assessment plan, he focused exclusively on the assessment task and only made one, somewhat vague, reference to the task in connection with the curricular unit: "I think this would be an ending assessment." However, in the third assessment plan, Dean situated the entire collection of assessment tasks within a curricular unit, indicated by phrases such as "In a unit on convex & concave lenses," "I would start with…," "I would follow up the lab," and "At this point." Not only did Dean indicate the placement of assessment tasks, but also indicated the purposes: "These Q's [questions] are meant to discover commonalities in S [student] thinking to guide formative assessment, & to lead Ss [students] to consider key ideas." Moreover, he demonstrated a full cycle of inquiry by considering student commonalities (what students know?), a learning goal ("lead to Ss to consider key ideas"), and strategies to get students there (guide formative assessment). He was the only teacher who demonstrated a full cycle of inquiry on an assessment plan.

In summary, the Use dimension explored the teachers' expertise at using assessment to support students' science learning. The teachers did in fact shift from a view that assessment serves primarily an evaluative role to serving a formative role to support students' science learning. The changes presented reflect mostly discrete strategies that could be situated within instruction, rather than expertise in planning an entire cycle of formative assessment focused centered on specific learning objective, as the case with Dean.

Equity Dimension: Acknowledging the Role of Language While Assessing

By considering how the teachers provide opportunities for students, regardless of English language proficiency, to demonstrate scientific knowledge and access a rigorous science curriculum through assessment, the Equity dimension explored expertise in equitably assessing ELs. One overarching pattern emerged during the teacher interviews: The teachers *became more aware of the role of language while assessing*. However, this understanding only translated to a certain extent while planning and critiquing assessment.

Assessment Understanding

The first interview revealed that while cognizant that they *should* be focusing on equity issues, the teachers came into the program with a range of tentative views about equitable assessing. One teacher confessed that she had no idea about what it means to equitable assess science, while some indicated that equitably assessing means being fair and unbiased, and yet others focused on multiple forms of assessment and knowing where all students are in the learning process. The teachers' views changed in that more of them acknowledged the integration of language and science as an important aspect of equity, even in assessment, and that the teacher is responsible for providing support that can help diverse students.

The teachers who recognized bias issues from the onset often talked about how students come into class with various cultural and academic backgrounds, and because of that, should be assessed with a variety of assessment forms. These

teachers changed in that they began focusing on specific language demands that ELs would encounter during assessment, like Glenda:

> The quizzes [by my cooperating teacher] were done orally…So I've been thinking about that because students have a hard time listening, like English language learners. So it's hard for them [ELs] to listen and be like trying to look at their word bank…I've been thinking like well the oral part takes away their reading of the question, but then also it's a different aspect that you're like testing I guess. You're listening and whether to not they can understand spoken English well. (Interview 3)

Glenda recognized that listening, accessing and using a word bank, and reading the question are all language demands that ELs would have to navigate on her cooperating teacher's quiz.

On the other hand, teachers who could articulate some strategies for addressing language on assessment at the onset of the program became more aware about the specific role of language in assessment *and* how to scaffold language. Scaffolding draws on sociocultural approaches based originally on the work of Vygotsky (1978) and consists of pedagogies that determine, and facilitate, what students can do on their own versus what they can do with a more capable peer. At the onset, the strategies discussed by the teachers attempted to circumvent language: "I will explain a question in English, and then I will try to draw a really good picture…so even if you can't totally understand the language, you can…see the diagram and you can be able to figure out what is going on" (Dean, Interview 1). By the end of the program, these teachers explicated knowledge ways to modify language, although they often did not identify how to move students to higher levels of learning: "besides the technical like vocab[ulary] we're learning that's related to

the content…[I would] keep the language at their level" (Lauren, Interview 3). Other scaffolds articulated by the teachers include graphic organizers, modeling instruction, setting expectations (often through the rubric), and science notebooks. Modeling, in particular, figured prominently into the teachers' understanding of ways of setting expectations about what content was being assessed and help students to master that content:

> I think it [my assessment practices] changed to the point where I need to model what I want to use as assessment a lot more. They have to be able to see what is coming so that they can do better at it. It's harder to ask them to think at a higher level, in fact, you have to say that I am asking you a higher level question here, this is not going to be easy and I want you to try. (Darlene, Interview 3)

Finally, it is not exactly clear to what extent the teachers came to understand that they should be assessing language use. Some teachers, such as Dean, embraced the integration of language and science and considered "arguments [and] other kinds of academic language…as a vehicle for…addressing the actual concept" (Interview 3). However, even Dean implied difficulty in supporting, rather than eliminating students' use of language while assessing:

> I mean the goal is to make language less of an issue…The goal is to…try to understand what they [students] know about content without … docking them for language … I feel like it's more valid assessment is what I am trying to say… of their knowledge of the concept.

Although he valued the integration of language and science to make a "more valid assessment," Dean felt the need to curtail the influence of language, which would make it difficult to support fully ELs' language use through assessment.

Like the previous dimensions, the teachers' demonstrated changes in their understanding of equitably science assessment. Yet, compared to the previous dimensions, there was less change in facility with equitable science assessment.

Assessment Facility

While teachers increased in their awareness of the role of language while assessing, this awareness did not fully translate to changes in their facility to consider language while planning assessment. Similar to the range of views on equitable science teaching and assessment, the teachers' assessment plan scores reflected a greater initial range in facility than either the Construction or Use dimension (see Figure 6). On the second and third assessment plan, the fairness scores converged onto level two expertise, with the exception of Lauren and Dean, who remained stable in their scores across all three plans. Between assessment plan two and three, the teachers fluctuated between demonstrating level one and level two expertise on the access sub-scale. With the exception of Dean, no teacher scored higher than a two on either sub-scale.

On the third assessment plan, the teachers typically included some implicit attention to sociocultural influences, often through the use of multiple assessment forms:

> I would have them do a lab in which they work to find whether something is an acid or base & mystery ones to figure out → I could do more inquiry and let students pick from some acid/base experiments and they could do a presentation to the class or they could do a research project about a certain acid or base or whatever else. I would give a quiz with vocabulary, and diagrams for students to show me what they know making sure students are understanding with each assessment!

In this plan, Glenda provides students with multiple ways to demonstrate scientific understanding – via presentations, research projects, and multi-part quizzes. However, she did not describe ways in which she made the assessment fairer for ELs by considering the language of assessment. Similar to other teachers, none of her assessment tasks solicited students' *use* of language in science explicitly – although it is possible through presentations and research projects. Those teachers that did consider the use of language in assessment did so in the form of having student write scientific explanations or respond to open-ended questions related to scientific explanations. However, no teachers made connections between scientific discourse and EL learning in the assessment plan, which represents a level three expertise in the access sub-scale.

Referring back to broad patterns, descriptively over time, the teachers' scores to both their assessment plan and assessment critique increased in the Equity dimension. Yet, the teachers generally scored higher on their assessment critiques. Qualitatively, this score differential was reflected by identifying specific strategies that could scaffold language use, predominately modeling:

> I think it was effective to first model explanations verbally with the group. It may have been helpful to also have those explanations on the board and to show how they break down into "claim" and "evidence" to discuss the form. (Dean, Assessment critique 3)

The teachers drew even moreso upon their awareness of the role of language on the *Learning Theories* final project and PACT commentary. The teachers often incorporated at least one specific strategy for modifying or scaffolding language in assessment:

> Looking at the assessment now, I realize that adding pictures to the worksheet would greatly benefit ELL and LEP [limited English proficient] students. Another idea to keep in mind while assessing groups is making sure to not grading strictly on presentation skills. Some ELL and LEP students may not feel comfortable talking in front of the class. Asking students to write down what parts of the group work they helped on would make for a fairer grade. (Hallie, *Learning Theories* final project)

The previous three examples highlight differences in the teachers' facility with equitable science assessment. Overall, they were more attentive to language in assessment in the situated assessment plans and assessment critiques than in the decontexualized assessment plans. Michael's comments during his second interview may shed some light on these differences:

> My knowledge has changed about what an equitable assessment looks like …but not how to make one. Someone can give me an assessment and I can say that this is definitely equitable and I can point out where we can see if there are some areas that might not be so equitable because we have had practice but creating it from ground up, I do not feel confident about that. (Interview 2)

To summarize, when looking closer at assessment expertise via interviews, open-ended survey responses, and teacher products, many patterns emerged. Both quantitative and qualitative analyses reveal differential patterns across the three conceptual dimensions. There also appears to be differences in assessment understanding versus assessment facility. The next section discusses the importance of these changes in light of what the literature says about how teachers should assess science.

Discussion: Smooth Sailing and Rough Waters While Developing Assessment Expertise

Science classroom assessment is a complex process. To become an expert at it entails knowing about a range of assessment forms and functions. It means using assessment in the context of instruction and considering the needs of diverse learners. Yet, while first learning about assessment among other areas of teaching, how easy is it for preservice teachers to develop expertise in science classroom assessment? The changes described in this paper suggest aspects of assessment expertise that the teachers were able to understand and apply (smooth sailing) and those aspects that that did not develop as smoothly (the rough waters).

Before discussing the findings, some limitations should be noted. First, to caution, the small sample size (N = 11) and exploratory nature of the instruments limit generalization of quantitative results to larger populations. Instead, quantitative data from surveys provide broad patterns within the sample to triangulate with more in depth qualitative data. As an example of this limitation, on average across all three likert scales and on each administration, the teachers scored above a 3 out of a possible 4 (between *agree* to *strongly agree*). Such high scores, even from the onset, suggest a possible ceiling effect or leading questions. The scores still showed significant change across two dimensions; however, those changes may be greater if the survey was more sensitive to change.

The interpretive features of this study also present challenges. As Putman and Borko (2000) caution, "Rather than pretending to be objective observers, we must be

careful to consider our role in influencing and shaping the phenomena we study" (p. 13). I have described the structure and goals of the assessment-focused instruction to communicate what the teachers were exposed to and my involvement in the process. Although I hypothesize that the assessment-focused instruction played a significant role in their developing expertise, the purpose of this study was not to test the effect of the intervention on assessment expertise, but rather to describe the changes that occur when exposed to such instruction. The interaction with the participants may not be appropriate for a controlled experimental design. Yet, by allowing me to understand the context surrounding teacher learning and by allowing the participants to become familiar with my presence while they invite me into their thinking, the interaction was appropriate for the study's exploratory and descriptive nature.

Smooth Sailing: From Evaluation to Instructional Support

The Wilcoxon-signed rank test, descriptive analysis of score distributions, and qualitative analyses collectively provide substantial evidence that the teachers' smoothest sailing was shifting toward a formative view of assessment, a view necessary for assessment-capable teachers (Abell & Siegel, 2011; Black & Wiliam, 1998b; Czerniak & Lumpe, 1996; Hill, Cowie, Gilmore, & Smith, 2010). There are several possible explanations for this shift. The teachers initially associated assessment with "tests" and explicated limited knowledge of formative assessment (Graham, 2005). However, early on the teachers were exposed to formative assessment views and strategies through the assessment-focused instruction. Considerable research in science education has theorized about and investigated

assessment through a formative lens and this well developed conception of assessment has likely translated into teacher education programs. In fact, previous research indicates that through teacher education programs, preservice teachers do change in their knowledge about formative views (Buck, Trauth-Nare, and Kaftan, 2010; Graham, 2005; Yilmaz-Tuman, 2008).

Across all dimensions, the teachers also moved toward a more situated view of assessment – connecting assessment to what is supposed to be learned and the instructional unit. This shift echoes teaching expertise literature in that teachers shift from knowing *about* content, pedagogy, and student characteristics (e.g., being an EL) to knowing *how* to apply knowledge in particular contexts (Bransford, Brown, & Cocking, 2000; Dreyfus & Dreyfus, 1986). Whereas at the onset of the program, the teachers planned assessment and talked about assessment during the interviews in terms of the isolated tasks – what to assess, what form it would take – by the end of the program they consistently considered when assessment would occur within a curricular unit and for what purpose. They best demonstrated this situated nature of assessment on the PACT commentary. PACT's emphasis on a formative use of assessment may help explain their demonstrated expertise on it. Rather than expect teachers to explicate the technical quality of assessments, the PACT rubrics were associated with feedback quality, steps to modify instruction, monitoring student learning, and other features congruent with a formative perspective.

Through a situated perspective of cognition in teacher learning, Putnam and Borko (2000) argue that teacher learning should be grounded in actual teaching

practices. The PACT and features of the assessment-focused instructed not only provided the teachers with a context in which to apply assessment understandings, but also gave them a space in which to reflect on their assessment beliefs and practices. These spaces have been shown to improve teachers' assessment knowledge and practices (Briscoe & Wells, 2002; Buck, Trauth-Nare, & Kaftan, 2010; Sato and Atkin, 2007), exemplified by Whitney's response during the second interview:

> The first time we interviewed I just kind of had these vague ideas about assessment and now I have some much more tangible evidence about assessment. So that's probably the biggest change. Just seeing…how just with the assignments and with how they [cooperating teachers] run things in class and adapting lessons as opposed to reading something about formative versus summative assessment. I can actually see the different types in action and how students respond to it.

Rough Waters: From Strategies to Frameworks

Consistent with a study by Siegel & Wissehr (2011), the teachers encountered rougher waters while moving from assessment understanding to the realities of planning assessment, particularly equitable science assessment. By examining changes over time and interpreting teacher interviews, this study provides a more nuanced look at such struggles.

The Wilcoxon-signed rank test, descriptive statistics, and qualitative analysis all point to differences between teachers' facility with equitable science assessment and other assessment dimension. Over time, the teachers explicated more awareness about language issues in assessment and knowledge about ways to scaffolding language in assessment. However, the teachers rarely applied such knowledge in their assessment plans. Although research has advanced our understanding about using

assessment formatively, research on equitable classroom assessment is still in its infancy and may not be well translated into teacher preparation. Therefore, beyond the assessment-focused instruction, the teachers may have received little additional support to develop their expertise in the Equity dimension.

Siegel and Wissehr (2011) suggest that teachers' beliefs about learning and assessment values, not just their assessment knowledge, may contribute to the divide between theory and reality. Well formed, and informed, beliefs aid teachers in flexibly applying assessment principles while planning assessment, rather than readily accepting theoretical orientations. The teachers expressed some beliefs, particularly around the formative use of assessment, echoing science education and assessment reform. Yet, by analyzing varying levels of expertise, the findings indicate that the waters may have become rougher for the teachers as they *integrated* beliefs and knowledge for the purposes of acting upon them in practice.

The teachers understood the importance of aligning tasks to a particular learning objective and considered the learning objective while planning assessment. However, they did not consider the full coherence among the task, learning objective, and evaluative criteria. Also related to the Construction dimension, the teachers often expressed higher capacity to modify assessments already designed, rather than construct assessments from the ground up. I had explicitly taught the assessment triangle to the students and used it as a guiding framework, but the teachers never referred to it while planning or while critiquing assessment. A richer sense of theory guiding particular learning objectives and assessment tasks may also help them apply

assessment knowledge and beliefs while engaging the entire process of assessing science and coherently linking assessment components (Bol & Strange, 1998; Lyon, 2011).

In a similar vein, the teachers explicated knowledge of formative assessment strategies and drew upon some of those strategies in the assessment plan. Yet, on the assessment plan, few teachers incorporated an entire cycle of cycle of inquiry in which they addressed (a) the learning objective, (b) what patterns or what conceptions they would look for in student responses, *and* (c) how they would modify instruction based on what they learn. Only by considering all components would the teachers demonstrate, fully, a formative use of assessment as outlined in the literature.

In the Equity dimension, the teachers primarily changed by adopting an awareness of equity issues (e.g., role of language) and by explicating knowledge of some strategies for making science assessment fairer for ELs. While teachers, such as Hallie, often professed a capacity to "tweak [the language of] assessments," and thereby assess more fairly, they professed having difficulty designing assessment them from the onset in ways that will be fair. Similar to the other dimensions, the teachers did not fully draw upon a cohesive framework. In *Science Education Theory* course assignments, some teachers referred to Siegel's (2007) "McCes" framework for equitable classroom assessments, read earlier in the course; however, none referred to it in interviews or in assessment plans.

To summarize, the teachers smoothly shifted from an evaluative to formative view of assessment, a shift reflected in their assessment plans. They also began

situating assessment within instruction. However, the waters became rougher with other dimensions of assessment expertise, particularly equitable science assessment. Furthermore, it was rougher to reach the highest levels of expertise, which called for them to integrate assessment understandings into a more cohesive framework.

Contributions and Next Steps

This study contributes to science education conceptually, methodologically, and empirically – contributions that set the stage for future research and for preparing preservice science teachers.

Conceptually, the CUE framework enhances previous conceptualizations of assessment expertise by including the Equity dimension. Linguistically diverse classrooms are no longer isolated to particular geographic regions in the United States, and more and more teacher education programs are expecting teachers at all grade levels and disciplines to teach and assess in ways equitable for all students. A second conceptual contribution is moving from "lists" of what science teachers needs to know and learn about science assessment to "levels" of assessment expertise. The patterns identified early in analysis were used to refine theoretical levels of expertise, in a recursive manner. For instance, a 4-point expertise rubric was used in this study to evaluate teacher responses to hypothetical scenarios.

Methodologically, this study used methods consistent with previous research examining teacher beliefs and thinking. Tensions will naturally arise while exploring changes in expertise across an entire teacher education program, instead of just patterns within a single science methods course. For one, the survey and interview

used in this study to compare changes across time were not situated within the teachers' actual program and, thus, limited in authenticity to program tenets. On the other hand, data sources such as the PACT, well situated within the program, only occurred once and thus could not be used to track changes over time. Even classroom observation, not reported in this paper, became problematic since the teachers' level of participation varied drastically from the beginning toward the end of the program. Despite the drawbacks, the combination of interviews, surveys, and more ecologically valid program artifacts, such as the PACT commentary, collectively solicited various aspects of assessment expertise and were analyzed in various ways to describe broad and more in depth patterns.

The empirical contribution is rooted in this study's exploratory and descriptive nature – suggesting patterns of change that may appear in other samples of teachers. Two overarching patterns should be noted the hold the most promise for directing future research and teacher preparation. First, the patterns of change varied by conceptual dimension. Second, by looking at a small set of preservice teachers through varying levels of expertise, this study provides a more complete and in depth description of assessment expertise changes, both quantitatively and qualitatively. For example, while the evidence demonstrates that the teachers were able to appropriate and apply formative understandings while planning assessment, the pattern differed from their capacity to equitably assess science in assessment plans. Both empirical contributions are important as science education moves from theorizing about how teachers should assess to putting research into practice.

The conceptual, methodological, and empirical contributions can inform future research. To move in the direction of researching assessment expertise of a large sample of science teachers, several important next steps need to happen. First, researchers need to validate instruments and protocols, both quantitative and qualitative, which could apply to a variety of studies. For instance, the survey scales in this study might be measuring a narrow scope of each dimension or a construct only somewhat related to the dimension. In a similar vein, the open-ended prompts might underestimate assessment facility. For example, the teachers generally scored higher in the *Learning Theories* final project and PACT commentary, most likely because of the expanded guidance, attachment of the assignment to course content and evaluation, as well as longer span of time given to complete and receive feedback on the task.

Using instruments validated with a larger pool, researchers can compare samples of preservice teachers from various teacher education programs, different conditions (e.g., instructional intervention), and to experienced teachers to determine more generalized models of assessment expertise growth. The patterns of change identified can also lead to refined coding schemes in studies with a larger pool of science teachers to investigate why changes do or do not happen. Continued research that unpacks patterns and looks into influences on those changes will provide a better account of barriers and successes as science teachers become prepared to assess. For instance, using classroom observation data, I described how three of the teachers (Dean, Glenda, and Lauren), all with different expertise trajectories, grew in their

awareness of language's role in assessment and how they treated language while assessing in their teaching practicum (see Chapter 3).

The conclusions drawn from this study can also inform the preparation of future science teachers. As a personal experience, based on the findings, I restructured how I sequenced instruction during the following cohort's *Learning Theories* course, where I took on an expanded role as a section lecturer. In particular, I focused on the alignment between task and learning objective first, knowing that this would be an important first step before considering evaluative criteria that might align with the task and objectives.

The findings suggest that preservice secondary science teachers, even when exposed to instructional focusing on assessing linguistically diverse students, need more opportunities to enact such understandings. Teacher educators can draw on preservice teachers' experiences in the classrooms, both classroom observations and student teaching, to reflect on observed disparities between ELs and non ELs, language demands of science assessment that may be challenging for ELs, as well as ways in which language demands may afford ELs with opportunities to use and develop language. Considerable attention should be given to incorporate and scaffold authentic literacy tasks in science assessment so that all students, but particularly ELs, have access to scientific discourse. To support teacher preparation, instructional materials for science teacher educators can be developed, and tested, that emphasize science assessment in linguistically diverse classrooms. Finally, evident by the differential patterns of change across dimensions, it would benefit science teacher

educators and science teachers themselves to consider assessment in light of the multiple dimensions. While discussing and providing assessment-related instruction, teacher educators can make explicit references to the multiple assessment conceptions and purposes, particularly a focus on equity and assessment. Practicing science teachers can reflect on their own expertise and consider how to navigate the rough waters.

The differential changes among the conceptual dimensions also warrant further attention to what aspects of assessment are emphasized in teaching standards. For example, California's Commission on Teacher Credentialing (2009) laid out several teaching standards associated with assessing student learning – mostly with a formative use of assessment – and addressing academic needs for diverse students. However, the only direct connection between assessment and ELs is situated within a single sub-standard on monitoring instruction, phrased as "*use assessment results to plan instruction to support English learners"* (p. 14). The findings from this study suggest that developing expertise at equitable science assessment is a greater struggle than developing expertise at using assessment formatively. Therefore, teaching standards, which guide teacher preparation and support, may benefit by including standards on the assessment of linguistically diverse students.

To conclude, over the last two decades, science education researchers have advanced our understanding about how diverse students learn and represent scientific knowledge, have developed tasks aimed to elicit complex scientific thinking and practices, and have studied the relationship between science learning, language, and

assessment – collectively leading to arguments and theoretical positions about *how* teachers should assess science. However, science education cannot overlook the realities of appropriating the assessment knowledge and beliefs called for by such positions and applying this understanding. Articulating the extent and nature of change will help researchers explore the *process* of developing assessment expertise and identify influences on developing expertise. The hope is that with continued research, not only will researchers understand how teachers should assess science in diverse classrooms, but also provide evidence that will inform how preservice science teachers are prepared in this arduous and vital endeavor.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer Netherlands.

Aguirre, J., & Speer, N. M. (1999). Examining the relationship between beliefs and goals in teacher practice. *The Journal of Mathematical Behavior, 18*(3), 327-356.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ash, D., & Levitt, K. (2003). Working within the zone of proximal development: Formative assessment as professional development. *Journal of Science Teacher Education, 14*(1), 23-48.

Aydeniz, M. (2006). *Understanding the challenges to the implementation of assessment reform in science classrooms: A case study of science teachers' conceptions and practices of assessment*. Unpublished doctoral dissertation, The Florida State University, Tallahassee.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education.* Dordrecht: Kluwer Academic Publishers.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Buckingham: Open Univ Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Bol, L., & Strage, A. (1998). The contradiction between teachers' instructional goals and their assessment practices in high school biology courses. *Science Education, 80*(2), 145-163.

Bransford, J., Brown, A. L. & Cocking, R. (2000). *How people learn.* Washington, DC: National Academy Press.

Briscoe, C., & Wells, E. (2002). Reforming primary science assessment practices: A case study of one teacher's professional development through action research. *Science Education, 86*(3), 417-435.

Buck, G. A., Trauth-Nare, A., & Kaftan, J. (2010). Making formative assessment discernable to pre-service teachers of science. *Journal of Research in Science Teaching, 47*(4), 402-421.

Campbell, C., & Evans, J. A. (2000). Investigation of preservice teachers' classroom assessment practices during student teaching. *The Journal of Educational Research, 93*(6), 350-355.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought process. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255-296). New York, NY: Macmillan.

Commission on Teacher Credentialing. (2009). *California standards for the teaching profession*. Retrieved from http://www.ctc.ca.gov/educator-prep/standards/CSTP-2009.pdf.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3-21.

Creswell, J. W., & Plano Clark, V. L. (2007). *Design and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Czerniak, C. M., & Lumpe, A. T. (1996). Relationship between teacher beliefs and science education reform. *Journal of Science Teacher Education, 7*(4), 247-266.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120-138.

Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York, NY: The Free Press.

Furtak, E. M. (2009). *Formative assessment for secondary science teachers*. Thousand Oaks, CA: Corwin Press.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., & Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment, 11*(3&4), 237-263.

Genc, E. (2005). *Development and validation of an instrument to evaluate science teachers' assessment belief and practices*. Unpublished doctoral dissertation. Florida State University, Tallahassee.

George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference, 11.0 update (4th Ed*). Boston, MA: Allyn & Bacon.

Gerard, L. F., Spitulnik, M., & Linn, M. C. (2010). Teacher use of evidence to customize inquiry science instruction [Electronic version]. *Journal of Research in Science Teaching*, 47(9), 1037-1063.

Gipps, C. V., & Murphy, P. (1994). *A fair test?: Assessment, achievement and equity.* England: Open University Press.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through preservice teacher education. *Teaching and Teacher Education, 21*(6), 607-621.

Harlen, W. (2003). *Enhancing inquiry through formative assessment*. San Francisco, CA: Exploratorium.

Hesse-Biber, S., Dupuis, P., & Kinder, T. S. (1991). Hyper RESEARCH: A computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis. *Qualitative Sociology, 14*(4), 289-306.

Hill, M., Cowie, B., Gilmore, A., & Smith, L. F. (2010). Preparing assessment-capable teachers: What should preservice teachers know and be able to do?. *Assessment Matters, 2*, 43-64.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Jones, M. G., & Carter, G. (2007). Science teacher attitudes and beliefs. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1067-1104). Mahwah, NJ: Lawrence Erlbaum.

Kagan, D. M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks principle. *Review of Educational Research, 60*(3), 419-469.

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education, 90*(5), 820-851.

Lee, O. (2004). Teacher change in beliefs and practices in science and literacy instruction with English language learners. *Journal of Research in Science Teaching, 41*(1), 65-93.

Luft, J. A. (1999). Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education, 10*(2), 107-121.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

Lyon, E. G. (2011). Beliefs, practices, and reflection: Exploring a science teacher's classroom assessment through the assessment triangle model. *Journal of Science Teacher Education, 22*(5), 417-435.

Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education, 21*(2), 133-149.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

McMillan, J. H. (2007). *Formative classroom assessment: Theory into practice.* New York, NY: Teachers College, Colombia University.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future*. London: King's College London.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained. *Journal of Mixed Methods Research, 1*(1), 48-76.

Morrison, J. A., & Lederman, N. G. (2003). Science teachers' diagnosis and understanding of students' preconceptions. *Science Education, 87*(6), 849-867.

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

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Otero, V. K., & Nathan, M. J. (2008). Preservice elementary teachers' views of their students' prior knowledge of science. *Journal of Research in Science Teaching, 45*(4), 497-523.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Pease-Alvarez, L. & Hakuta, K. (1992). Enriching our views of bilingualism and bilingual education. *Educational Researcher*, 21(2), 4-19.

Popham, W. J. (2006). *Assessment for educational leaders.* Boston, MA*:* Allyn & Bacon.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education, 27*(2), 472-482.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.

Saldana, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Sato, M., & Atkin, J. M. (2006). Supporting change in classroom assessment. *Educational Leadership, 64*(4), 76-79.

Sato, M., Coffey, J., & Moorthy, S. (2005). Two teachers making assessment for learning their own. *Curriculum Journal, 16*(2), 177-191.

Schemp, P. G., Manross, D., Tan, S. K. S., & Fincher, M. D. (1998). Subject expertise and teacher's knowledge. *Journal of Teaching in Physical Education, 17*, 342-356.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences (3rd Ed)*. Boston, MA: Allyn and Bacon.

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2003). *Evaluating new approaches to assessing learning.* Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295-314.

Shaw, J. M., Bunch, G. C., & Geaney, E. R. (2010). Analyzing language demands facing English learners on science performance assessments: The Sald framework. *Journal of Research in Science Teaching, 47*(8), 909-928.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational Measurement (*4th Ed*).* Westport, CT: Praeger Pub Text.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching, 44*(6), 864-881.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391.

Smith, M. U., & Siegel, H. (2004). Knowing, believing, and understanding: What goals for science education? *Science and Education, 13*(6), 553-582.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3-21). New York, NY: Routledge.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-573.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan, 83,* 758-765.

Stoddart, T., Bravo, M. A., Solís, J. L., Mosqueda, E., & Rodriguez, A. (2011). Effective Science Teaching for English Language Learners (ESTELL): Measuring pre-service teacher practices. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA. April 2011.

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

Stoddart, T., Solis, J., Tolbert, S., & Bravo, M. (2010). Effective Science Teaching for English Language Learners (ESTELL). In D. Sunal, & C. Sunal (Eds.), *Teaching science with Hispanic ELLs in K-16 classrooms* (pp. 151-181). Albany, NY: Information Age Publishing.

Tomanek, D., Talanquer, V., & Novodvorsky, I. (2008). What do science teachers consider when selecting formative assessment tasks? *Journal of Research in Science Teaching, 45*(10), 1113-1130.

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. *Science Education, 85*(2), 137-157.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Webb, N. (2002). Assessment literacy in a standards-based urban education setting. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design. Alexandria, VA:* Association for Supervision and Curriculum Development.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Yilmaz-Tuzun, O. (2008). Preservice elementary teachers' beliefs about science teaching. *Journal of Science Teacher Education, 19*(2), 183-204.

Yung, B. H. W. (2006). *Assessment reform in science: Fairness and fear.* New York, NY: Springer Publishing Company.

Footnotes

[1]For example, while being asked about the purpose of assessment a teacher may respond with "to diagnose learning" (professed) although it is unclear whether he or she just know about the diagnostic role of assessment or they personally accept such a role. Given the difficulties of separating knowledge from beliefs, this study combines both constructs into a single one – assessment understanding.

[2]A sociocultural perspective views learning as occurring through the social interaction between teacher and students and between students (Tharp and Gallimore, 1988; Vygotsky, 1978). Assessment plays a role in the learning process, and therefore, is also part of these social interactions. The content being assessed represents the type of knowledge valued in education (Gipps, 1999) and provides a means by which students and teachers communicate with each other about progress toward learning goals. Through a sociocultural perspective, such progress aligns with Vygotsky's (1986) notion of the zone of proximal development (ZPD), or the region between what students can do on their own and what they can do with a more capable peer.

[3]Hacienda University and all teacher names are pseudonyms.

[4]The PACT served as useful data source for several reasons. Upon examination of the rubric criteria and tasks, I found general alignment between the PACT and dimensions of the CUE Assessment Expertise Framework. In particular, the PACT's emphasis on academic language and EL teaching/learning throughout the PACT aligned well with the equity dimension. It was also well validated, evident by the PACT Consortium's extensive measures to ensure the technical quality of the assessment (Pecheone & Chung, (2007). Finally, similar to the *Learning Theories* Final Project, I did not provide feedback on nor score the Teachers' final products. Instead, the teacher supervisors provided Teachers with feedback on drafts, prepared them for the tasks, and scored the PACTs using a set of rubrics for each task.

Appendix A

Secondary Science Assessment Survey – Likert scales

Construction scale
1. It is important to consider validity while constructing assessments of student learning (the extent to which the assessment measures what it is supposed to measure).
2. It is important to consider reliable scoring while constructing assessments of student learning (the extent to which different scorers give the same score).
3. Teachers should make use of a variety of assessment strategies to determine what students know & are able to do.
4. Assessments should draw on theories of how students learn science.
5. Assessments should align with the desired learning objectives
6. Any criteria/rubric used to interpret student work should match the desired learning objectives.
7. It is important to consider how the language demands of an assessment (e.g., length of questions/prompts, vocabulary, types of grouping structures) might affect inferences made about what students know and can do.
Use scale
1. Teachers should monitor student learning by eliciting student responses that require scientific thinking and discourse.
2[a]. Appropriate assessment feedback makes little contribution to student learning.
3. The information from assessment should be used to modify future instruction.
4. Teachers should elicit student understanding throughout a science lesson to guide the rest of the lesson.
5. Student assessment of their own work can help them regulate their own science learning.
6. Students should be involved in designing assessments in science classrooms.
Equity scale
1. Assessing students' scientific thinking and reasoning contributes to student access to a quality science education.
2. Assessment should inform students what is important to learn in science.
3. Assessments should elicit both productive (speaking/writing) and receptive (listening/reading) modalities.
4. While constructing assessments of student learning, it is important to consider how the assessments will benefit individual or groups of students.
5. The act of assessing shapes students' motivation, confidence, and knowledge of what is important to learn in science.
6. The format/type of tasks used to assess student learning should resemble the type of tasks used throughout instruction.
7[a]. Criteria used to assess student learning should not be shown to students before they are assessed on them.
8. While assessing student learning, the students' linguistic and cultural background should be considered.
9. While assessing student learning, teachers should include appropriate accommodations for students with limited proficiency in English.

[a] Indicates an item that is negatively worded

Appendix B
Vignette from Secondary Science Assessment Survey – Assessment Critique Prompt

Ms. Sanchez wants her 9th grade Biology students to use evidence as they explain laboratory investigation findings. She assumes that her students have never engaged in evidence-based explanations before; therefore, after a laboratory investigation on osmosis, she models different ways to use evidence, such as "I claim that…" and "the following evidence….led me to the claim…" Then, in groups of four, students orally practice explaining their findings.

Later on in the unit, Ms. Sanchez provides each student with information and data about another investigation related to osmosis. She then asks students to "write an explanation about what is happening in the investigation based on the information and data provided." Students individually complete the task and turn their response into Ms. Sanchez. She uses a rubric to score the results, which contributes to the students overall class grade. She informs each student of his/her score and keeps the actual student responses.

Appendix C
Teacher Interview Prompts (without probes)

Teacher Assessment Interview 1

1. First off, when you hear the word "assessment" what are the first words or phrases that come to mind?
2. Could you please describe your experience being assessed in science classrooms? K-12, undergraduate, or graduate school.
3. Could you please describe any experience you have had learning about educational assessment.
4. How do you think students effectively learn science?
5. What does it mean to equitably teach science?
6. How would you describe to a fellow science teacher what it means to assess student learning?
7. Hypothetically, you are asked to construct an assessment of student learning. What are some things you would consider when constructing it? Why?
8. What would you do with the assessment information you gathered about the students? Why?
9. I'm going to show you the prompt and your response to one of the open-ended survey items you answered last week. *[show prompt and response]* Can you take me through the response again and explain your reasoning for the aspects you thought were effective and ineffective?
10. Finally, what does it mean to you to equitably assess student learning?

Teacher Assessment Interview 2

1. Can you describe your experience so far throughout the teacher education program courses learning about assessment
2. How have your cooperating teachers assessed student learning?
3. Have your cooperating teachers explicitly discussed opinions or strategies about assessing student learning?
4. Can you describe your experiences assessing student learning in your teaching placement?

PACT Reflection – part of Teacher Assessment Interview 3

Now, I am going to ask specific questions about the focal PACT assessment – that is, the task you used to analyze student work [show or describe task].

9. Can you please take me through the structure of the assessment, what it assessed, and why you chose it.
10. Do you think that all of your students had a fair chance to show what they knew or could do on the assessment? Why or why not?
11. How did you know whether your students learned the learning objectives being assessed?
12. Do you think the assessment contributed to student learning about [the learning objective]?

CHAPTER 3

WHAT ABOUT LANGUAGE?: PRESERVICE TEACHERS' EVOLVING EXPERTISE IN EQUITABLE SCIENCE ASSESSMENT

Abstract

The study I describe here was part of a larger research plan designed to explore the ways in which secondary science preservice teachers' assessment expertise, which included equitable assessment for English learners, changed during their 12-month long teacher education program. Qualitative analyses revealed that part of moving toward equitable assessment involves becoming increasingly aware of the role of language while assessing science (see Chapter 2). Through interviews, responses to hypothetical assessment scenarios, and observation of classroom practice, I explore the theme of language in equitable science assessment further by describing the evolving expertise of three preservice teachers – Dean, Glenda, and Lauren. The case studies reveal how teachers' awareness toward language in assessment was reflected in terms of using multiple forms of assessment, particularly scientific discourse, to uncover students' conceptual understanding and by identifying the language demands inherent in assessment. I discuss the findings in terms of two tensions present in the literature and which the teachers experienced: whether they should be assessing language as well as scientific understanding *and* to what extent they should be reducing the language demands of science assessment versus scaffolding them. The value of this paper lies in its descriptive account of the struggles and successes of being preparing to assess science in linguistically diverse classrooms, a critical line of research.

Introduction

In terms of political, monetary, and psychometric challenges, Lee and Luykx (2006) have suggested that "Valid and equitable assessment of nonmainstream students remains one of the thorniest difficulties in educational policy and practice" (p. 100). Although their statement was in reference to large-scale assessments given at the state and national level, it would not be a stretch to make the same claim for classroom assessment.

To start, equity, as conceptualized in this paper, is rooted in the notion of *opportunity to learn*. Thus, an equitable classroom is one in which a student's race, ethnicity, sex, socioeconomic status, or language proficiency does not preclude him or her from accessing challenging content instruction, in the case of this study, science. This does not mean every student should learn in the same way. Instead, to teach equitably means recognizing how particular groups of students are typically underserved in the educational systems and ways to support student learning for particular groups of students. This conception of equity extends to assessment in that all students also need appropriate opportunities to *demonstrate* what they have learned.

Assessment plays a critical role in the learning process. Teachers assess to make decisions about student advancement through secondary science coursework, influencing what they learn, as well as make decisions about teaching and student learning through the information gathered by assessing. There is a growing body of research that examines inequitable opportunities for English learners (i.e., students

who are still developing English proficiency) to demonstrate what they know and can do in content area instruction, particularly math and science. This study aims to contribute to such literature on ELs and assessment through the context of preparing secondary science teachers to assess science equitably in linguistically diverse classrooms.

The number of non-native English speakers in K-12 classrooms across the United States is significant, with an estimated 11.2 million (or 21%) of 5- to 17-year-olds in 2009 (National Center for Educational Statistics [NCES], 2011). Yet, teachers and teacher education programs are not fully prepared to meet the linguistic and academic demands of EL students (Bartolomé, 2002; Wong-Fillmore, 2007). We also know from past research that preservice teachers can enter teacher education programs with deficit beliefs about nonmainstream students – viewing them as less capable, not being aware of cultural and linguistic influences, and not viewing multicultural teaching (including teaching language) as their responsibility (Bryan and Atwater, 2002). Such beliefs can have profound effects on how preservice teachers conceptualize equitable science assessment and how they assess science in diverse classrooms.

Scholars also claim that teachers are not being adequately prepared to assess in ways that aim to support, instead of just evaluate, student learning (Shepard, 2006; Stiggins, 2002). While critical to science education research, more work is needed to describe whether teachers are being prepared to assess science equitably, as well as to

identify struggles or tensions that may arise while being prepared to equitably assess science.

I had previously reported on a mixed methods analysis of 11 secondary science preservice teachers (hereon referred to as "teachers") moving through a 12-month long teacher education program at Hacienda[1] University (see Chapter 2). The teachers were all part of a cohort that took classes together, including a quarterly seminar course led by two teacher supervisors. With cooperation of faculty members and the supervisors in the program, I led instruction to the teachers focused on assessing linguistically diverse students. The intent of the larger analysis was to explore their changing expertise in science classroom assessment. In this paper, I chronicle three of the teachers – Dean, Glenda, and Lauren – to provide a more descriptive account of one pattern that emerged – *the teachers' increased awareness about the role of language in assessment.*

Equitable Science Assessment: Conceptual and Theoretical Foundations

Sociocultural View of Assessment

Given its complexity and various perspectives, equity must be appreciated fully through multiple dimensions, as opposed to a single definition (Seceda, 2008). The same reigns true for conceptualizing equitable *assessment*. As Stobart (2008) states, "Fairness [in assessment] is fundamentally a sociocultural, rather than a technical, issue" (p. 346). Instead of attending solely to systematic bias in assessment, teachers should attend to the unique needs of ELs while assessing science. ELs are more than just second language learners, whose language proficiency influences how

they demonstrate scientific knowledge. ELs also bring with them perspectives and experiences shaped by their home, language, and culture, which collectively influence how they and others *perceive* their ability to do science. Therefore, the foundation of conceptualizing equitable science assessment must have its roots deeply embedded in sociocultural theory. The problem of equitable assessment, therefore, is socially, culturally and historically based. In order to uncover the assumptions, biases and working practices that inform equitable assessment, we must first examine assessment through a sociocultural perspective.

In essential ways, a sociocultural perspective sheds light on the relationship between classroom assessment and science learning. First, science learning is enhanced when instruction and assessment occurs in contexts that are culturally, linguistically, and cognitively meaningful to students (Fusco & Barton, 2001; Lee and Fradd, 1998; Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001). Next, assessment is a tool in which students and teachers necessarily communicate with each other about progress toward learning goals, particularly when teachers use assessment formatively. Through a sociocultural perspective, formative assessment works in Vygotsky's (1986) notion of the zone of proximal development (ZPD), or the region between what students can do on their own and what they can do with a more capable peer (see also Wells, 1999). Furthermore, facilitated by assessment, the social interactions between teacher and students and between students, inform students what type of knowledge is valued in education and how to represent this knowledge (Gipps, 1999). These social interactions promote learning ideally when

part of joint, dialogic activity (Tharp and Gallimore, 1988). Finally, just like the learning process, while assessing via these social interactions, language serves as a mediating tool (Vygotsky, 1978).

The Role of Language in Equitable Science Assessment for ELs

Attention to language figures prominently in promoting educational excellence and equity, because language (words, ideas, and reasoning) can allow students to participate in classroom activity; thereby accessing the rigorous subject matter valued by the community (La Celle-Peterson & Rivera, 1994). Yet, ELs come into science classrooms without the fully developed academic language needed to participate in science. Specific to science learning, Kelly (2007) suggests, "Researchers have approached issues of equity from a language point of view" to "understand the ways in which students get access to knowledge, and to consider the knowledge that counts as science in given circumstances" (p. 455). Guided by a focus on language, to assess science equitably means to provide opportunities for all students, regardless of English language proficiency, to demonstrate what they know and can do in science (*fairness*) and, through assessment, provide students with access to a rigorous and relevant science curriculum (*access*). I expand on these two points next.

Fairness. Teachers provide opportunities for all students to demonstrate what they know and can do in science by first recognizing that in any content-area assessment a student's proficiency in the language of the assessment is also being evaluated (American Educational Research Association [AERA], American

Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Teachers can focus on modifying language features of the assessment task itself, since vocabulary, sentence structure, and discourse patterns all influence how students perform, particularly ELs (Abedi & Lord, 2001; Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Solano-Flores & Nelson-Barber, 2001). However, teachers must be careful not to deny students with access to literacy tasks, including reading and writing science, by attempting to reduce the challenging features of language. Instead, teachers can focus on scaffolding, meaning that they engage in pedagogies that establish and work within students' ZPDs (Walqui & van Lier, 2010). Teachers can scaffold language in assessment by modeling language use, providing visual support, and making the assessment content relevant to student culture and the local and the physical environment (Siegel, 2007; Stoddart, Solis, Tolbert, & Bravo, 2010). By scaffolding, instead of merely reducing, language in assessment, teachers can better address sociocultural influences by drawing on them as resources instead of factors that need to be controlled (Solano-Flores, 2010).

Access. Focusing on language in assessment can address inequitable conditions by providing ELs with access to a rigorous science curriculum, even when they may be still acquiring the academic language possessed by their non-EL counterparts. Although such access is important for all students, ELs are historically relegated to less rigorous science courses, perhaps due to perceptions about their limited ability to engage in complex thinking (Pease-Alvarez and Hakuta, 1992;

Valdes, 2001) Thus, teachers can engage in equitable science assessment, for one, by promoting students' regular use of language in science classrooms. Researchers have found that activities in which teachers integrate scientific inquiry and language use are particularly supportive for EL learning (Stoddart, Pinal, Latzke, & Canaday, 2002). Such activities can form the basis for science assessment as well. For example, in science performance assessments, teachers observe student performance or observe an authentic student product (e.g., a lab report). In these assessments, ELs have been found to productively use the language of science while they demonstrate inquiry abilities, such as writing scientific arguments (Lyon, Bunch, & Shaw, 2012).

Research Objective

The conceptual and theoretical grounding of equitable science assessment just explicated encouraged me to analyze multiple sources of qualitative data so that I could determine, when enrolled in a teacher education program emphasizing equity/language issues and exposed to assessment-focused instruction emphasizing equity/language issues,

1. In what ways do the three teachers evolve in how they understand equitable science assessment and apply this understanding while planning and critiquing assessment?
2. In what ways is this evolution reflected or not reflected in how the teachers assess science during a culminating teaching event?

Study Context

The location and curricular emphasis of the teacher education program at Hacienda University provides an ideal context in which to study the teachers as they become prepared to assess science in linguistically diverse classrooms. Hacienda University is located in the state of California, neighboring communities that house a culturally and linguistically diverse student population – the most common non-English native language being Spanish. The teachers complete their student teaching in these diverse schools. As part of the program, all teachers obtain a Crosscultural, Language, and Academic Development (CLAD) certificate, allowing them to provide instruction for English Language Development (ELD) and Specially Designed Academic Instruction Delivered in English (SDAIE). Several of the program's courses emphasize teaching for culturally and linguistically diverse students, such as *Social Foundations of Education, Teaching and Learning in a Diverse Society*, and *Methods of English Language Development.*

During the study, I led assessment-focused instruction for the teachers via three courses – *Teaching and Learning in a Diverse Society* (Summer 2010), *Science Education Theory* (Fall 2010), and *Science Methods* (Winter 2011). Instruction emphasized equitable science assessment in several ways. For instance, the science teachers took *Teaching and Learning in a Diverse Society* along with other secondary preservice teachers. In one 2-hour class devoted to assessment, I discussed the different functions of assessment and specific ways in which the culture and language of assessment might be unfair for ELs. The science teachers then read and discussed Solano-Flores & Nelson-Barber's (2001) article on cultural validity of science

assessment in a separate discussion section to consider deeply the influence of language and culture while students demonstrate *scientific* knowledge. In *Science Education Theory*, the teachers engaged in four assessment case studies, where they read about, watched, or participated in various science assessments (e.g., performance assessment), discussed the scenario through guided prompts (some of which focusing on equitable science assessment), then individually wrote how they would modify the assessment to make it more equitable for English learners. Finally, in *Science Methods*, the teachers planned an assessment to be used in their student teaching and analyzed student data, again considering a linguistically diverse student population during the planning and student analysis.

Toward the end of the program, the teachers completed the Performance Assessment for California Teachers, or PACT, a culminating teaching event used in over thirty California universities to determine whether teachers have demonstrated proficient competencies associated with the California teaching standards. To complete the PACT, teachers planned and implemented approximately a week's worth of lessons, videotaped two self-chosen segments of these lessons, and wrote an extensive commentary, usually between 20 and 30 singles spaced pages. In the commentary, they articulated how they planned instruction and assessment, analyzed student work, and reflected on the effectiveness of their teaching. Throughout the commentary, the teachers were expected to address teaching and assessing for linguistically diverse students and include lessons plans, artifacts, and sample student work.

Approaching Case Studies: Method, Selection, and Analysis

Analysis in the Larger Study

I employed a mixed methods methodology in the larger research plan to track measurable changes of the teachers' assessment expertise over time in the program and to describe qualitative changes that occurred. As indicated in Table 1, the data sources included interviews (see prompts in Appendix A), likert scale and open-ended survey items (see Appendix B), and various teachers products completed for the program (see Table 2). Furthermore, I observed two lessons from the teachers' PACT teaching event – seven to eight months after the onset of the program (February to March 2011) – to capture how they applied their assessment expertise in classroom practice. I audiotaped each observation and wrote field notes about what was being taught (i.e., the learning objective) and the various classroom activities, including how the teacher was teaching the content (what participant structures and tasks) and assessing student learning, informally and formally. To understand the teachers' own accounts of the activities, I collected their written PACT commentaries and asked the teachers to reflect on how they assessed student learning during the PACT, including how they addressed issues of equitable science assessment.

Qualitative data were analyzed through iterative rounds of coding to identify longitudinally patterns (Saldana, 2009). In terms of quantitative data, teacher responses to the two open-ended survey items, an *assessment plan* and *assessment critique*, were scored, with two other trained scorers, using a rubric (see Chapter 1) in three different assessment expertise dimensions, one of which associated with

Equitable science assessment. Finally, likert scale items were grouped into three reliable scales, one of which also associated with equitable science assessment. Both likert item scales and scored open-ended responses were analyzed descriptively for changes over time (mean response, distribution of responses) and tested for statistically significant changes from beginning to the end of the program through a Wilcoxon signed rank test (analogous to a paired sample *t*-test).

The results of the mixed methods analysis (see Chapter 2 for more details) indicated that, on average, the teachers did change in terms of their expertise at equitable science assessment, although the quantitative and qualitative analyses described these changes in different ways. For instance, the teachers' average response to *Equitable science assessment* likert scale items changed significantly (according to a Wilcoxon signed rank test), whereas the assessment plan and assessment critique scores did not. The qualitative analysis of written content revealed a host of patterns, including the focal topic of this paper, that the teachers *became more aware about the role of language while assessing*. This theme, combined with the finding that only their beliefs toward equitable science assessment, but not their facility with equitable science assessment, changed through quantitative analyses, set the stage for more in depth research.

Table 1

Data Sources across the Year

Beginning of the program (July 2010)	During the program (August 2010 to April 2011)	Toward the end of the program (May 2011)	Notes

Interview 1	Interview 2 (December, 2010)	Interview 3	Semi-structured with question probes (see Appendix A)
Assessment plan and critique 1	Assessment plan and critique 2 (December, 2010)	Assessment plan and critique 3	Written responses to two open-ended prompts (see Appendix B)
	Teacher products	Teacher products	See table 2
		PACT observations	Audiotaped. Included written field notes and self-reported reflection during interview 3

Table 2

Written Teacher Products

Teacher Product (when collected)	Description
Teaching and Learning in a Diverse Society final project (August 2010)	Write a description of three activities to teach a particular science standard and two assessments to assess the learning objectives. In the description, identify (a) relevant learning theories, (b) language demands, and (c) responsiveness for diverse learners
Science Education Theory assessment case studies (October – December 2010)	Based upon the assessment scenario you observed/participated in and discussed, what modifications would you make (if any) to the assessment and how the information from the assessment is used? How would these modifications support learning and promote equitable assessment for ELs?
Science Education Theory equity essay (November 2010)	Using specific examples from class discussions and readings to date, address the following issues: 1) What do you see as the major equity/diversity themes in science education? 2) What does it mean to contextualize science instruction and what is the rationale/purpose for contextualization? 3) What does it mean to equitably assess students in science and what is the rationale/purpose for this?
Science Education Theory final project	Based on your research into the central concepts, facts, procedures, beliefs, and connections for this topic, describe one appropriate strategy for assessing your topic. Specially

(December 2010)	discuss: 1) What theories of learning does the assessment task connect to? 2) How will you interpret what students know and can do? 3) How can the assessment be used to support learning and other goals associated with that topic? 4) How will you address issues of equity, particularly for English learners?
Performance Assessment for California Teachers (May 2011)	Task 1: Context for learning Task 2: Planning instruction and assessment Task 3: Instructing students and supporting learning Task 4: Assessing student learning Task 5: Reflecting on teaching and learning

The Researcher

At the time of the study, I (the author) was a Ph.D. Candidate using the data gathered as part of my dissertation research. I had previously taught high school science in culturally and linguistically diverse classrooms across California, including an integrated science class for intermediate ELs. My first contact with the participants came as I introduced myself to them at the onset of their program, a few days before the first interview. The teachers were aware of the purpose of my study, to understand how their expertise in assessment changed, and knew about the two roles I would play throughout the year – as someone researching them, as a participant observer, and as a guest instructor helping to support their learning about science classroom assessment. Through my role as a guest instructor, I was able to converse informally with teachers, often around issues of assessing science. Although ethnographic data collection was not officially part of my study, and thus not analyzed systematically, such conversations naturally influenced my interpretations about their thoughts, struggles, and successes.

Case Study Teacher Selection

I was interested in looking closer at three teachers from the original 11that represented a range of trajectories in terms of their expertise at equitable science assessment. I examined two measures analyzed in the larger study – changes in an aggregated likert scale score and changes in an open-ended survey item (their assessment plan) score (see Table 3). Based on the measurable changes, I selected Dean since he was consistently at or above the average score, but only improved on the scaled belief score, whereas Glenda was also above average, but only improved on the assessment plan score. I selected Lauren, since she started out below the average score and improved in both scores, showing the greatest improvement of any teacher in the belief score (from a 3 to a 4).

Table 3

Case Study Teachers' Assessment Expertise Scores

	"Equitable assessment" scaled belief score (out of 4)		"Equitable assessment" scaled assessment plan score (out of 8)	
	July 2010	May 2011	July 2010	May2011
Dean	3.57	3.67	5	5
Glenda	3.57	3.56	3	4
Lauren	3.00	4.00	2	3
Cohort average	3.29	3.66	3.09	3.64

Case Study Analysis

I used a case study methodology to describe Dean, Glenda, and Lauren's evolving expertise in equitable science assessment to "retain the holistic and meaningful characteristics of real-life events" (Yin, 2009, p 4). I compiled all relevant qualitative data for Dean, Glenda, and Lauren and reanalyzed those data with

attention to their understanding of and facility with equitable science assessment. In particular, I located text that indicated how Dean, Glenda, and Lauren conceptualized equitable science teaching and assessment, their beliefs toward and knowledge about equitable assessment and related issues, and knowledge of strategies to address the role of language in assessment (either in terms of fairness or promoting access to science). Drawing on the multiple sources of data, I wrote narratives in a chronological structure to describe each teacher's (a) expertise at equitable science assessment expertise at the beginning of the program, (b) evolving expertise over the span of the program, and (c) equitable science assessment in practice. The narratives aim to convey, in a persuasive manner, representative changes for each teacher. I conducted member checks (Guba and Lincoln 1989) with Dean, Glenda, and Lauren to enhance the trustworthiness and validity of the reported narratives. The teachers did not need me to make any changes.

Preservice Teacher Case Studies: The Evolution of Expertise in Equitable Science Assessment

Dean - "Assessment in Discourse"

Dean is a 26 year old White male, who is a native English speaker with beginning second language proficiency in French. He holds a B.S. in Physics and has tutored undergraduate students in Physics prior to entering the teacher education program.

Equitable science assessment in the beginning. Coming into the program, Dean conceptualized equitable science teaching as not being biased against particular

groups of students and as focusing on student *progress*, as opposed to just achievement. Dean's emphases on bias and progress carried over to his conceptualization of equitable science *assessment* in that he believed teachers should use multiple assessment forms to draw on students' various strengths; thereby not privileging one learning style. Dean expressed some awareness that language influenced how students performed while they were assessed in science. For Dean, the solution to addressing language issues was to "be as visual as possible [while assessing] so you [the students] don't really even need language to understand the problem." Thus, Dean's goal was to "bypass the whole language thing." Dean stated that one way to bypass language was for students to respond in their native language, which Dean thought could be plausible if he actually spoke the students' native language proficiently.

Dean incorporated several forms of assessment during his initial assessment plan, focusing on students' ability to reason scientifically about reflecting and refracting light. However, he did not offer any strategies for making the assessment fairer for ELs nor address any issues of language in assessment. In a similar vein, while critiquing Ms. Sanchez's assessment practices on the hypothetical scenario, Dean thought that her written assessment of evidence-based explanations was "helpful for understanding a students [*sic*] understanding of theory or ability to utilize deductive reasoning," although he would have liked more focus on mechanistic explanations.

Evolving expertise. As he progressed through the program, Dean recognized that equitable science teaching and assessing included the identification of language demands (i.e., what students have *to do* with language). He acknowledged that language becomes a barrier for students, particularly ELs, to access science content. Instead of believing that teachers should use multiple assessment forms to account for learners' varying strengths and learning styles, Dean thought that teachers should use multiple assessment forms to take into account any *language demands* associated with assessment. Dean demonstrated knowledge of specific language demands that could be challenging for ELs, such as complex text on rubrics. He also discussed more strategies that would mitigate the "negative" influence of language – making succinct rubrics and valuing students' everyday English language while formatively assessing students with the intent to expect, eventually, curriculum-based vocabulary.

Dean also believed that focusing on scientific discourse in the classroom allowed the teacher to integrate science and language development by "letting language and science sort of build on each other because they are kind of one in the same for science." Toward the end of the program, Dean expressed a more specific account of how to integrate science and language through discourse – by promoting argumentation and explanation in the classroom. According to him, arguments and other kinds of academic language are "a vehicle for….addressing the actual concept." Through reflection of discourse in his teaching practice, he increasingly realized the importance of incorporating multiple forms of assessment: "I saw so many students unable to put an argument altogether in writing…but then, be perfectly capable of

demonstrating knowledge…in conversation." Overall, Dean had developed a position, coined by him, of "assessment in discourse," meaning that he thought that he could best uncover student thinking by engaging them individually in dialogue around the concept of interest. However, although he used discourse (a form of language) as a way to find out what students knew and could do, he still was uncertain about whether he should *be assessing* language use in addition to science content:

> It's going to be hard to sort of separate, um, assessing the language versus assessing the content, and I think it's my job to teach them language, but I guess I'm unclear as to whether I should be grading language improvement on top of content improvement or understanding.

On Dean's second and last assessment plans, instead of just describing multiple assessment forms, he explicitly mentioned that he would assess in multiple ways. An even bigger change was his attention to language. He still asked students to *write* scientific explanations, but also stated that the writing might "be scaffolded with a questionnaire," so that "assessment would be limited with regards to grammar & syntax but extensive with respect to students [*sic*] grasp of content." In a similar vein, on the third assessment plan he mentioned multiple ways to assess and made an explicit reference to scaffolding language and science in assessment:

> I would follow up the lab with a think-pair-share based on the previous questions. The "share" is an informal yet formative assessment in that I get to hear how Ss [students] think about the key ideas, and at the same time scaffold content and language through my responses. At this point, a quick quiz where Ss [students] respond by completing a diagram could be administered to check for understanding in a more visual and less language-dependent fashion.

While critiquing the assessment scenarios, Dean's responses changed in that he addressed ways to support student understanding of evidence-based explanations. For instance, on the second assessment critique, Dean believed that the rubric used by Ms. Sanchez "may not yield the type of specific feedback needed to sufficiently inform students of the strengths and weaknesses of their evidence-based explanation." On the last critique, he expanded on the rubric's limitations by specifying how the rubric could be more useful for students:

> I think that using a rubric is helpful, but Ss would likely benefit more if they had seen the rubric beforehand, and were given a marked rubric afterwards. This would make expectations transparent…If the S got a graded rubric AND comments there would be more scaffolding for the S [student].

On the last assessment critique, Dean also wanted to know the specific English proficiency of the students and suggested that Ms. Sanchez model the structure of explanations by breaking them into "claim" and "evidence" for students.

Equitable science assessment in practice. Dean's PACT teaching event occurred within a high school conceptual physics class, described by Dean as the lowest (in academic rigor) of the three physics offered at the school, with the others being college-prep and advanced placement. All 28 students in the class were 11th graders, and 13 (46.4%) of them were English learners. The central focus of the learning segment was that *charges exert forces on each other and that those forces are responsible for the way charges move*. Although Dean identified evidence-based explanations as one of the national science standards, his specific learning objectives focused on conceptual, including mechanistic, understanding instead of explanations as a language learning objective.

In the PACT commentary, teachers were expected to write out their theoretical stance. For Dean, language figured prominently: "A belief in the interdependence of language and thinking (Vygotsky, 1986) lies at the core of my instructional design for developing my students' knowledge and abilities in both science and academic language" (PACT, task 2). Dean used several assessments throughout the lesson segment, including concept checks, a lab, and a jigsaw poster presentation to "stimulate the development of academic language" (PACT, task 2). I focused my analysis on the latter assessment.

Related to a lab investigating electric charges, students completed the *Jigsaw Poster Presentation*, in groups of four, by drawing a diagram of electric charges and arrows to show the movement of charges, and by writing explanations on the poster. Students then presented their poster to Dean as he circulated among the groups, probing student understanding more deeply using open-ended questions.

To interpret what students knew and could do, Dean accompanied the jigsaw poster (and presentation) with a rubric divided into three dimensions – key elements, explanation, and participation. The *key elements* criterion focused on conceptual understanding, indicated by arrows pointed in the correct direction. The *explanation* criterion focused on how students explained *where* and *why* the electrons move, drawing on their understanding of force. Finally, the *participation* criterion focused on ensuring all members of the group worked together and that *all* members were able to explain the poster.

As part of the PACT, the teachers analyzed student work on two levels – by interpreting patterns of whole class performance and by interpreting individual work of three students, one of which was an EL. Overall, Dean was cognizant of performance differences between his ELs and non-ELs in terms of providing explanations:

> My English language learners…. maybe this is just the language barrier but they didn't…didn't provide explanations…some of them just didn't provide explanations at all um, for the quiz, … didn't really go…didn't really include the justification… despite my best intensions [*sic*], the English language learners really didn't do…as well as their…the fluent peers. (PACT reflection)

Dean offered one explanation for the disparity – that students "simply did not understand the directions" (PACT, task 5) and described a future modification to the assessment: "project the quiz on the board while I read the directions, and visually indicated the steps to which the directions refer" (PACT, task 5). I observed Dean asking students about the poster directions, indicating he may have been cognizant even during the assessment about the potential language demand of the directions. Dean also identified what his ELs *were* able to do:

> Given the explanations that many ELLs [ELs] were able to produce today with a little scaffolding, I feel that a more informal, discursive form of immediate assessment could make this an equitable grading practice for groups insofar as I am able to assess understanding of content. This is not to say that academic language is not something worth assessing, but it should be done on an individual basis and not be reflected in the groups' grade. (PACT, task 5)

Regarding feedback, Dean stated, "A recurring theme of both written and oral feedback was about helping students construct explanations that are specific and supported" (PACT, task 4). Oral feedback came immediately in the form of dialogue

with each student; however, I observed this feedback mostly in terms of their conceptual understanding, rather than use of language (i.e., structure of the explanation, use of vocabulary). In the PACT, Dean described several activities completed after the learning sequences based on the assessment information, much of this focused on students' explanation of the concepts. However, although Dean found a disparity between the explanations of his ELs versus non-ELs, the instructional modifications did not appear to target EL learning.

Glenda - "Excellence through Explanations"

Glenda is a 21 year old White female with no second language proficiency and a B.S. in Environmental Science. As an undergraduate, she participated in CalTeach, an undergraduate program offered at several California universities to recruit and begin preparing mathematics and science teachers. It is through the CalTeach program that Glenda was first exposed to the notion that students might not all do well on the same type of assessments, since they have different strengths and learning styles.

Equitable science assessment in the beginning. In the first interview, Glenda professed that she had difficulty describing what equitable science teaching means or looks like to her: "I have a hard time distinguishing between equal and equity. I know there's a difference, but I can't remember." However, she was more specific about equitable *assessment,* believing that it meant finding out what *all* of her students knew, despite the differences in learning styles – e.g., whether they best demonstrate understanding through writing, visually, or kinesthetically. Thus, to her a

critical aspect of equitable science assessment was providing multiple forms of assessment. In particular, Glenda believed that providing students opportunities to *explain* what they knew was effective "because [while] explaining something, you kind of have to really know what's going on…I feel like explaining it in person makes a student kind of have to understand it better." Glenda's reasoning extended to ELs. When asked about how she would assess science if she had ELs in her classroom, she responded with "maybe a…written like assessment so they could practice writing…Like science writing and writing about science, but then also like oral so I can see that they know what's going on even if they may not be able to write it out."

Glenda's focus on multiple assessment forms and explaining carried over into her initial assessment plan:

> I would have students be able to describe the different types of faults *and have students be able to draw them and show which plates are moving in which direction*. Having students *describe and be able to draw what is happening is more important* than them answering a multiple-choice question that doesn't show exactly what the student knows. (Assessment plan 1, emphasis added),

as well as her initial assessment critique:

> "Having them [students] orally practice is also good because *some students may be better at verbal explanation rather than written*. (Assessment critique 1, emphasis added)

In both examples, Glenda noted that multiple forms of assessment (describing *and* drawing; explaining verbally *and* through writing) should be used and that students should be given opportunities to explain.

Evolving expertise. Over the span of the program, Glenda continued to believe in incorporating multiple forms of assessment, including explanations, so that all students had opportunities to demonstrate what they knew. However, Glenda's understanding of equitable science assessment evolved in that she focused more so on language, rather than differences in learning styles, in relation to both ELs and students with disabilities:

> Thinking about it [assessment] in terms of like what would be equitable for those types of students [ELs and students with learning diabilities], it's like put in more perspective…A student who has a motor or writing disability like really can't do this assessment, like I never thought of that!

For example, while describing how she would modify the assessments encountered in the four *Science Education Theory* assessment case studies, Glenda consistently referred to how explanations in the assessment could support EL learning:

> Asking students to explain their answers and use complete sentence[s], I feel that EL students would be able to learn English better and are able to use it more accuratley [*sic*] in a science context. Having the students explain their answers I could understand what they are thinking and why they formed their opinions and could give more accurate feedback on their responses and work. (Assessment case study 2)

During the third interview, while she again considered multiple forms of assessments while assessing, Glenda also considered specific language demands of assessment, such as "what kind of reading goes into the questions or the instructions." In particular, Glenda was concerned that how her cooperating teacher, who administered quizzes orally, might not be equitable for ELs because "it's hard for them to listen and be like trying to look at their word bank." Therefore, Glenda suggested adding written prompts to the oral directions. As put by Glenda, "Is it…equitable for all

students if you have them reading or listening or a bunch of writing? Are all students gonna be able to do that?" (Interview 3).

Glenda's enactment of equitable science assessment evolved as well. First, she acted upon her understanding of writing scientific explanations in science assessments. While on the first assessment critique Glenda thought that it was effective for Ms. Sanchez to have students write explanations "because it can assess the students[‘] critical thinking/reasoning," on the second assessment critique, she thought it was important so that students could "practice using/saying academic answers…[and] because many students may not have much experience with this [writing/discussing explanations] and it is good for all students to be exposed to" (assessment critique 2).

Just as Glenda placed more importance on "using/saying academic answers," she also emphasized vocabulary use in assessment, which she did not attend to at the onset of the program. On the second assessment plan, Glenda asserted that as part of her criteria for assessing students' understanding of earthquakes, she would see if students could define associated vocabulary. Moreover, on the third assessment plan Glenda indicated that she would "give a quiz with vocabulary, and diagrams for students to show me what they know."

As expressed throughout this paper, addressing language in assessment means more than focusing on vocabulary. Glenda was able to incorporate more strategies to support ELs' use of language, beyond vocabulary, on assessment. For example, on the third assessment case study related to assessing written scientific arguments, she

described how she would use word glossaries and graphic organizers "to show the evidence, give their claim, and state limitations and assumptions that would work instead of writing a lengthy paragraph."

To summarize, then, Glenda's expertise in equitable science assessment, conveyed through her focus on multiple assessment forms and scientific explanations, evolved in that she considered explanations as not just a way to uncover what students know about science, but also a way to promote academic excellence for her students, regardless of English language proficiency.

Equitable science assessment in practice. Glenda completed her second teaching placement in a public middle school with only .09% ELs, the lowest population among the three schools in this paper. Glenda taught a sixth grade earth science class with 31 students, including only one EL and one former EL. However, 11 (35.5%) of the students had an Individualized Education Program (IEP) or were granted accommodations through a 504 plan. The central focus of the learning segment was for students to "become familiar with density calculations, understand what density is, and tying their knowledge of the rock cycle with density" (PACT, task 2).Glenda's content learning objectives focused on the transfer of scientific knowledge to novel situations. Additionally, she included one language objective: "Students will demonstrate that they can use their data to form conclusions about the lab using proper sentence form and punctuation."

Glenda's emerging focus on vocabulary was evident during classroom observation. On the first of the two observations, Glenda started class by presenting

students with two words, *bioaccumulate* and *concentration*, accompanied by definitions for students to copy into their notes. She elaborated on each vocabulary term and related the terms back to previous activities. After the vocabulary exposure, Glenda handed out a "pre-test," which she told the students would be used to see what they already know about density. Consistent with her belief in multiple assessment forms, Glenda gave students the opportunity to write or draw something. For example, she worded one question as followed: "Please describe what 'mass' means. Feel free to draw a picture if that will make your description clearer." On the second observation, the vocabulary of the day was *density*, which she again connected to previous activities and terms learned. She probed student understanding about density, and held up a rock to discuss density as it relates to rock types – a central focus of the focal assessment.

The focal assessment task was an investigation titled *How Compressed is a Rock?* Again consistent with her belief in multiple forms of assessment, students demonstrated scientific understanding by labeling diagrams, drawing, responding to short answer prompts, recording data, making calculations, and graphing. Moreover, Glenda recognized the benefit of having students work collaboratively and engaging in literacy tasks:

> If they [students] didn't understand exactly which column things [rocks] needed to go in, they could work, ask their group of students who could all help them and work as a little their own little community in their group. Um, I think that's more helpful than having them read and answer questions [individually] 'cause that doesn't really give them anything. It's like they really need to be able to talk…just them talking in their groups can extend their thinking because somebody else brought up an idea that they didn't even think of. (PACT reflection)

Glenda found learning differences between her mainstream students and nonmainstream students –ELs and students with IEPs. She reflected on what she could have done differently to support them:

> I have noticed that students with IEPs and English Language Learners tend to want more examples shown of what type of work is expected. I only gave one example, and this may not have been enough resources for students to be able to do well on this section.

In addition, throughout the PACT commentary and interview, Glenda reflected on how she could include sentence starters and more open-ended questions to scaffold language in the focal assessment. These scaffolds reflect her awareness of language in assessment.

Lauren – "Meeting the Language Needs of ELs"

Lauren is a 25 year old female, who has no second language proficiency. She holds a B.S. in Environmental Science and has no reported experience learning about educational assessment prior to entering the teacher education program.

Equitable science assessment in the beginning. At the onset of the program, Lauren associated equitable science teaching with ensuring that *all* students are learning, which to her meant not making assumptions about what individuals already know about science and meant relating the science curriculum to her diverse students. Lauren did not espouse a single definition for equitable science assessment, but rather associated it with a host of issues and practices, including objectivity and consistently while grading. Regarding ELs, she specified that equitable science assessment entailed providing different forms of assessment, since "it might be harder for them

[ELs] to respond in English to a complex science question." Furthermore, Lauren pointed out that she could "restructure the question in a way that ma[de] more sense to them" and could match the vocabulary used in assessment with the vocabulary used in instruction. Thus, Lauren was aware of some language issues, particularly for ELs, in her conception of equitable science assessment from the onset.

Lauren's initial assessment plan described how she would assess concepts related to Mendelian genetics by having students solve punnett square problems; however, she did not indicate how she would interpret student work (e.g., correct answer versus showing work), nor how she would address language issues while assessing. On the initial assessment critique, Lauren commented on two ways in which Ms. Sanchez did and did not support students' demonstration of scientific explanations. Lauren believed that Ms. Sanchez's use of groups would be supportive for students; however, Lauren believed that Ms. Sanchez should better align the content of the written assessment with the activities students had completed during instruction – giving students the opportunity to learn what was being assessed.

Evolving expertise. Early in the program, Lauren demonstrated specific ways in which to reduce language demands of assessment, such as by "rewrit[ing] all the test questions and answer choices in a non-verbose, simpler and straight forward fashion and omit unnecessary advanced academic language." She also believed that alternative forms of assessment, such as performance-based assessment would "give ELs a chance to show their knowledge without having to heavily rely on language" (Case study 2). Throughout the program, Lauren articulated, in greater detail,

strategies to *scaffold*, instead of reduce, language in assessment, such as by adding visual aids, providing the rubric at the beginning of the task to clarify expectations, and grouping students in ways that resembled instruction.

On top of becoming more knowledgeable about strategies to scaffold language in assessment, by the end of the program, Lauren expressed more concern for academic rigor in her assessment to meet the needs of her ELs. During the third interview, she insisted that instead of *omitting* advanced academic language on science assessments, she would keep "the language at their level – straight forward but not dumbing it down so it's really simple and complex because you know part of it is…Moving them up to the next level."

Although Lauren was knowledgeable of strategies that could scaffold language on assessment, she did not incorporate any of these strategies in her second assessment plan, and only one in her last assessment plan, related to group work in that students should "work together and…answer [questions] on their own. This way I know what they can do with help from peers and what they individually understand." On both the second and last assessment critique, Lauren continued to mention that Ms. Sanchez should assess in a way that matched instruction, but added that Ms. Sanchez should make expectations for the rubric more explicit by providing a rubric ahead of time. On both the plan and critique, she did not explain how nor explain why such strategies would specifically benefit ELs.

Equitable science assessment in practice. Like Glenda, Lauren completed her second teaching practicum in a public middle school. However, Lauren's school

housed a more ethnically and linguistically diverse student population with 51% of students identified as Latino and 23% ELs. Lauren's focal class had 30 students with nine (30%) ELs or former ELs.

Lauren's learning segment focused on "reproduction in flowering plants, with an emphasis on the complimentary [*sic*] nature of structure and function" (PACT, task 2). The numerous learning objectives (10 in total) in her PACT commentary included being able to "explain their [the students] scientific reasoning to answer lab questions." Thus, like Dean and Glenda, scientific discourse was a desired goal. However, unlike Dean and Glenda, Lauren categorized this focus on discourse as a language, instead of a science content, learning objective. Lauren's focal assessment was the *Flowers, Fruits, and Seeds Lab Report*, embedded within a multi-day investigation in which students observed and identified various flowers, fruits, and seeds. Lauren gave each student a worksheet with lab instructions as well as a set of questions that students completed as they completed the lab.

In several ways, Lauren applied her knowledge of supportive EL strategies during the focal assessment. First, instead of giving instructions for the entire lab and letting each group work at their own pace until completion, Lauren provided instructions for each section of the lab, let students work, and then facilitated/discussed questions related to that section of the lab, as represented in the field notes below:

> [At 28 minutes]. T [teacher] shows microscopic view of flowers and asks Ss [students] to look at carnations dipped in colored water and asks where on the flower most of the color is – refers them back to previous concepts (stomata).

> T instructs Ss to answer a question on data sheet (no right or wrong answer). I want to see inside of your head
>
> "Why do you think the petals are colorful, have a scent (smell), and have a unique shape? What could be their function in pollination?" (Field notes, Observation 3)

As indicated in the field notes, Lauren provided visual support by displaying the microscopic view of flowers to the class (though a doc-u-cam). She also included open-ended questions throughout the lab, instead of all at the end, and probed for student understanding of both the concepts and the lab instructions while circulating through the groups. She provided even more visual support and modeling of lab procedures through a PowerPoint slide show. Lauren explained that the step-by-step modeling allowed her to address the "language heavy" nature of the lab, which included "a lot of reading and listening and writing" (PACT reflection). As further support, Lauren provided one sentence frame to scaffold the students' explanation about seed dispersal strategies: "I think this seed is dispersed by the wind because ___."

Lauren interpreted student work by designing a rubric that focused on completing the questions and tasks, which she used for grading purposes. In addition, she constructed a rubric specifically aligned to each of the 10 *learning objectives*, articulated across three proficiency levels. Using the rubric as a guide, Lauren was able to reflect on student mastery of the two language objectives:

> Nearly all students had a good understanding of what counted as good evidence to support their predictions, but most students did not understand how to write their predictions and evidence in complete sentences and needed more scaffolding to achieve level 3 performance. On the other hand, the students who did write strong predictions supported by good evidence

> could have been further challenged by providing seeds whose dispersal strategy is less obvious.

Furthermore, she reflected on ways to *support* student mastery of the language objectives:

> Scaffolds I would provide include sentence frames for writing the seed dispersal predictions and evidence and the lab reflection. Another scaffold I have in mind is a matching exercise where students match written descriptions of each step in the process of flower reproduction from pollination to seed dispersal with pictures of each step. This would help them understand how to write about the process of flower reproduction[,] which directly addresses the essay question on the upcoming plant unit test. I would include more challenge questions in the lab where students applied what they learned or observed such as a question that asked students how they would classify a flower with 12 or 15 petals and also ask them to justify their classification.

Finally, as noted by Lauren,

> When teaching academic language to ELs, I need to be very explicit and provide opportunities to practice the academic language before it is required on assignments. This is suggested in the SIOP Model and evidence of the need for it can be seen in students' incomplete seed dispersal predictions.

In both example, Lauren's analysis of student work informed ways in which she could better meet their needs of her ELs in ways consistent with current research on effective learning for ELs.

Case Study Limitations

While steps were taken to ensure that the case studies accurately represented the teachers and followed rigorous qualitative methods, they should be interpreted in light of the particular context and the limitations of case study research. First, the teachers came from a program selected in part because it focused on equity/language issues and placed preservice teachers in linguistically diverse classrooms. Such exposure in itself may help the teachers develop a deeper understanding about the

role of language – they learn about it and experience it in the classroom. Although the purpose of this paper was not to describe how the assessment-focused instruction influenced their assessment expertise, the analysis was guided by the theoretical proposition that such instruction would be a key factor in their evolving expertise. Clearly, more work is needed to see how teachers from difference programs and different experiences consider equity in science assessment and, more importantly, see *what* influences their evolving expertise.

Second, although I drew upon and triangulated multiple sources of data, additional analyses, such as audiotaped or videotaped observations, could help elucidate the teachers' assessment expertise and provide more examples of equitable (or inequitable) science assessment in action. Lastly, the case studies are meant to provide rich, descriptive accounts representing a range of perspectives from the larger sample. These case studies are not meant to generalize to the larger populations of preservice science teachers. Instead, this exploratory study can lead to future analyses that test or explain changes in equitable science assessment. With this context and the limitations in mind, as discussed next, the cases contribute in important ways to conceptualizing and understanding equitable science assessment.

Discussion: What about Language while Assessing Science?

When discussing equitable science assessment, language is certainly a relevant topic. To ensure that all students, regardless of language proficiency, can fairly demonstrate academic content, teachers need to know about language and its relation to assessment (Trumbell & Solano-Flores, 2011). By giving ELs continued

access to rigorous science content and by furthering their language development, assessment can better promote equitable conditions (La Celle-Peterson & Rivera, 1994). In rich detail, the three cases demonstrate some of the ways in which secondary science preservice teachers can engage in equitable science assessment. The cases also demonstrate some of the tensions experienced by secondary science preservice teachers as they attempt to assess science equitably.

Dean, Glenda, and Lauren all evolved by viewing equitable science assessment as more than a technical issue (Stobbart, 2008). Most notably, they believed in using multiple forms of assessment to account for various language demands rather than just learning styles. Furthermore, they understood the importance of scientific discourse for ELs. However, they applied their understanding of equitable science assessment in various ways, not always fully congruent with their views.

Glenda began viewing assessment as more than just uncovering student understanding: "An equitable assessment is not only educational but is *meaningful*, and takes into consideration language and culture of the students" (Equity essay, emphasis added). Her focal PACT assessment, *How Compressed is a Rock?*, included multiple access points through drawing, graphing, and writing. Some questions prompted students to write explanations. She also thought that as students collaborated in a joint activity (Tharp & Gallimore, 1988), the discourse among them would help them "extend their thinking." However, throughout the assessment, Glenda did not appear to draw on students' home, culture, or local environment

(Stoddart, Solis, Tolbert, & Bravo, 2010), which could have made the assessment more *meaningful* to students. For instance, she could have situated the activity to the local geologic environment, allowing students to test rocks common to the area.

During the last interview, Lauren reflected on the importance scientific discourse for ELs in the form of argumentation:

> It's [scientific arguments] this other aspect of academic language that they [English learners] are trying to develop... And you know making a claim and making an argument and it's really important in science... But it's important in all aspects of life. I think so it's really important aspect of language that I think they need to be developing.

As part of the *Flowers, Fruits, and Seeds* investigation, her students wrote predictions and were supported in providing evidence through a sentence frame. Yet, the predictions, by Lauren's own admission, barely resembled the scientific arguments called for in science education. Furthermore, despite her emphasis on collaboration, there was little evidence that students engaged in argumentation while working in groups.

Finally, as Dean explained, "Structuring in opportunities for discourse is… beneficial for English language learners… to practice their language skills…[and] just practice talking, practice writing" (PACT reflection). During *the Jigsaw Poster Presentation*, he instructed students to explain electric charges through diagrams and writing. Unlike Lauren and Glenda, he also engaged each student in dialogue around the explanation. Thus, not only did he provide students with multiple access points, he applied what he had learned about integrating language and science, particularly

influenced by Stoddart et al.'s (2002) article about the importance of integrating inquiry and language to help ELs learn science.

Typically, scholars argue that argumentation promotes conceptual and epistemic understanding in science (Driver, Newton, & Osborne, 2000). In each case, the teachers came to view scientific discourse, such as argumentation, as also a way to promote language development. However, despite their evolving understanding about importance of language use, they did not fully apply their understanding, evident by their focus on conceptual understanding in the learning objectives and evaluative criteria. Next, I discuss two tensions that emerged as the teachers considered what to do with language while assessing science.

Tension 1: *Reduce* Language Demands of Assessment or *Scaffold* Them?

As conceptualized in this study, a key feature of equitable science assessment is to incorporate scientific discourse in assessment to give ELs access to rigorous science content. Early in the program, all three teachers wanted to reduce what students have to do with language in assessment, to "bypass the whole language thing," as put succinctly by Dean. However, as Trumbull and Solano-Flores (2011) remind us, "It is almost impossible to design a meaningful academic assessment that does not depend on language in some way" (p. 23). Science assessments that attempt to resemble authentic scientific practices will likely include a host of language demands that teachers must consider (Shaw, Bunch, & Geaney, 2010). Throughout the program, the teachers explicated knowledge of assessment accommodations, which, albeit with mixed evidence, could reduce bias against ELs. Some strategies

included simplifying the language, but not the content of test items, allowing students to use dictionaries or glossaries, and providing oral, in addition to written, instructions (Abedi, Hofsetter, & Lord, 2004).

We must also consider the opportunities afforded by language use while assessing, not just the challenges. Science discourse may amplify rather than simplify students' communication of conceptual understanding; thereby increasing science rigor (Walqui & van Lier, 2010). However, this opportunity is only realized if teachers scaffold students' use of language. As the year progressed, the teachers explicated more knowledge of strategies to scaffold language in assessment, not just reduce language. For instance, if listening to an oral quiz presented challenges for ELs, Glenda described how she could include the written questions for support. Dean proposed that engaging students in discourse while assessing would scaffold their scientific explanations. Lauren scaffolded language use in the *Flowers, Fruits, and Seeds* investigation and accompanying lab report by breaking down the assignment into manageable tasks that were discussed as a whole class before proceeding to each subsequent step. She also contextualized the investigation by bringing in local plants, provided extensive visual support, and engaged students in collaborative inquiry.

These practices afford ELs with opportunities to demonstrate what they knew via written scientific discourse, instead of assuming the students would not be able to write science due to the influence of language proficiency. However, knowing the linguistic challenges ELs might face on complex science assessments, it is difficult for teachers to expect students to communicate understanding through literacy tasks.

Lauren, in particular, had mixed feelings as to whether she should include science writing as part of science assessments due to fairness while grading students: "Depending on the level of English language proficiency... I would probably lean less on the writing. And now that I just mentioned... I really like the writing... I would probably provide some other type of [visual support]." While it appears that Lauren's instinct was to reduce the demand of language ("lean less on the writing"), she was able to explicate how to scaffold language use (through visual support).

While such language demands in complex assessments might be challenging for ELs, with appropriate scaffolding the demands might actually afford ELs with expanded opportunities to learn science and develop language (Lyon, Bunch, & Shaw, 2012). Formative assessment can play a particularly important role in scaffolding as students move through their ZPDs. Furthermore, scaffolding does not have to stop with conceptual understanding. Teachers also need to consider assessing students' use of language, for instance, by focusing on students' use of evidence in explanations and arguments, clarity of written or oral response, and use of vocabulary.

Tension 2: Assess Language Use *in addition to* Conceptual Understanding?

Another key feature of equitable science assessment is to use assessment as a way to further language development. Assessment can play a supportive role by helping teachers recognize what students are able to do with language and by using such information, provide instructional support that aids language development. As previously indicated, Dean, Glenda, and Lauren came to appreciate how ELs need to

be developing English proficiency as they learn science. They also incorporated some form of scientific discourse in their focal PACT assessment. However, they all professed or implied a tension as to whether they should be assessing language use *in addition to* conceptual understanding.

Although Dean believed that his job was to teach language and emphasized "assessment *in discourse*," he was not convinced that he should also be *assessing* discourse. Dean thought that probing for scientific explanations allowed him to interpret all students' conceptual understanding; thereby moving each student through his or her ZPD. He did not apply the same reasoning to assessing language use. Instead, he found the individualistic nature of assessing language problematic when trying to assign a group grade in the *Jigsaw Poster Presentation*. In essence, he had mixed feelings about giving a grade versus just giving feedback.

The teachers' rubric use during the focal PACT assessment also reflects the language use versus conceptual understanding tension. Rubrics could help teachers articulate those aspects of language they want to assess separate from content. Dean's rubric for the *Jigsaw Poster Presentation* included criteria related to the conceptual quality of the explanation: "*Student tells the direction of the forces on pith ball or electroscope leaves and describes force in terms of same or opposite charges.*" However, the rubric did not include criteria related to student use of evidence or curriculum-appropriate vocabulary – two criteria more congruent with language use. For the *"How Compressed is a Rock?"* lab report, Glenda's evaluative criteria also focused more on conceptual understanding – specifically, knowledge transfer –

instead of language use. Furthermore, the criteria did not reflect Glenda's vocabulary focus or her lone language objective, "Students will demonstrate that they can use their data to form conclusions about the lab *using proper sentence form, and punctuation*," (emphasis added). For both Dean and Glenda, a consequence of not assessing language was that the process of assessing lacked coherence. The learning objectives did not fully match the evaluative criteria, even though each teacher's focal assessment *could* have provided information on students' use of language.

Both tensions just described reify the complex interplay of assessing language and assessing content (Lee, Santau, & Maerten-Rivera, 2011). Without knowledge of students' use of language, it would be more difficult to support their language development through feedback or modifying instruction. This is why it is particularly important to understand what it means to equitably assess science and communicate that understanding to preservice science teachers. Although content and language are in reality, inextricably linked, for the purposes of supporting student learning, some criteria could focus more on language features while other criteria focus on conceptual understanding.

Concluding Remarks

Equitable science assessment is certainly a thorny issue for classroom practice, given the challenges both in ensuring ELs are able to demonstrate what they know and can do in science and, through the supportive role of assessment, ensuring that ELs access a challenging science curriculum. Without a deep understanding of equity issues while assessing science or opportunities to engage and reflect on

equitable assessment practices, science teachers might assess with only monolingual learners in mind. Alternatively, teachers might assume that ELs, due to limited English language proficiency, are incapable of learning *and demonstrating* complex scientific thinking – and thus deny them access to such knowledge. Therefore, it is of particular concern to understand how preservice science teacher garner expertise in equitable science assessment.

Dean, Glenda, and Lauren – who were all exposed to assessment-focused instructional centered on equitable science assessment – became increasingly aware of the role of language in assessment, and translated some of their understanding to classroom practice during the culminating PACT teaching event. However, they addressed the role of language differently – from Dean focusing on "assessment in discourse," to Lauren focusing more so on strategies to scaffold language. Overall, language was valued as a way for students to demonstrate scientific understandings, but less clearly applied as a way to promote equitable conditions.

While preparing teachers to assess science in linguistically diverse classrooms, *language*, and its relationship to assessment, is a necessary topic of discussion. In teacher education program courses, preservice teachers could analyze the language demands of various science assessments – from multiple-choice items to lab reports to performance assessments. They could integrate knowledge of science learning and language acquisition to strategize about ways to scaffold the language of assessment. Still, once teachers acquire some foundational knowledge about assessment, language, and equity, they may begin to wonder whether (and how) to

assess language use *in addition to* conceptual understanding and whether to reduce language demands of assessment or scaffold them. In other word, they may wonder, "What do I do with language?!" While the evidence from the case studies suggests that progress can be made, it will take considerably more effort to adequately prepare secondary science teachers to reconcile whether (and how) to assess for science equitably in actual classroom settings.

References

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1-28.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bartolome, L. I. (2002). Creating and equal playing field: Teachers as advocates, border crossers and cultural brokers. In Beykont, Z. F. (Ed.), *The power of culture: Teaching across language difference* (pp. 167-191). Cambridge, MA: Harvard Education Publishing Group.

Bryan, L. A., & Atwater, M. M. (2002). Teacher beliefs and cultural models: A challenge for science teacher preparation programs. *Science Education, 86*(6), 821-839.

Fusco, D., & Barton, A. C. (2001). Representing student achievements in science. *Journal of Research in Science Teaching, 38*(3), 337-354.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*(1), 355-392.

Guba, E., & Lincoln, Y. S. (1989). *Fourth generation evaluation.* Thousand Oaks, CA: Sage.

Kelly, G. J. (2007). Discourse in science classrooms. In S. K. Abell, & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 443-470). New York, NY: Routledge.

La Celle-Peterson, M. W., & Rivera., C. (1994). Is it real for all kids?: A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English-language. *Educational Researcher, 27*(4), 12-21.

Lee, O., & Luykx, A. (2006). *Science education and student diversity: Synthesis and research agenda*. England: Cambridge University Press.

Lee, O., Santau, A., & Maerten-Rivera, J. (2011). Science and literacy assessments with English language learners. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 254-274). New York, NY: Routledge.

Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education*. Advanced online publication. doi:10.1002/sce.21008

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

National Center for Education Statistics. (2009). *Science framework for the 2009 national assessments of educational progress*. Author.

Pease-Alvarez, L., & Hakuta, K. (1992). Enriching our views of bilingualism and bilingual education. *Educational Researcher, 21*(2),4-19.

Saldana, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Secada, W. G. (2008). Essay: What is equity in science education? In A. S. Roseberry, & B. Warren (Eds.), *Teaching science to English language learners* (pp. 167-182). Arlington, VA: NSTA Press.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4 Ed.). Westport, CT: Praeger Pub Text.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching, 44*(6), 864-881.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3-21). New York, NY: Routledge.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-573.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan, 83,* 758–765.

Stobart, G. (2008). Fairness in multicultural assessment systems. In H. Wynne (Ed.), *Student assessment and testing* (Vol. 4, Chapter 65, pp. 346-359). Thousand Oaks, CA: Sage.

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

Stoddart, T., Solis, J., Tolbert, S., & Bravo, M. (2010). Effective Science Teaching for English Language Learners (ESTELL). In D. Sunal & C. Sunal (Eds.), *Teaching science with Hispanic ELLs in K-16 classrooms* (pp. 151-181). Albany, NY: Information Age Publishing.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context*. England: Cambridge University Press.

Trumbell, E., & Solano-Flores, G. (2011). The role of language in assessment. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 22-46). New York, NY: Routledge.

Valdes, G. (2001). *Learning and not learning English: Latino students in American schools*. New York, NY: Teachers College Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: The MIT Press.

Walqui, A., & van Lier, L. (2010). *Scaffolding the academic success of adolescent English language learners.* San Francisco, CA: WestEd.

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

Wells, G. (1999). Dialogic inquiry: Toward a sociocultural practice and theory of education. England: Cambridge University Press.

Wong-Fillmore, L. (2007). English learners and mathematics learning: Language issues to consider. *Assessing Mathematical Proficiency*, 53, 333-344.

Yin, R. K. (2009). *Case study research: Design and methods.* Thousand Oaks, CA: Sage.

Footnotes

[1]The name of the program, teachers, and courses are pseudonyms

Appendix A
Teacher Interview Prompts (without probes)

Teacher Assessment Interview 1

1. First off, when you hear the word "assessment" what are the first words or phrases that come to mind?
2. Could you please describe your experience being assessed in science classrooms? K-12, undergraduate, or graduate school.
3. Could you please describe any experience you have had learning about educational assessment.
4. How do you think students effectively learn science?
5. What does it mean to equitably teach science?
6. How would you describe to a fellow science teacher what it means to assess student learning?
7. Hypothetically, you are asked to construct an assessment of student learning. What are some things you would consider when constructing it? Why?
8. What would you do with the assessment information you gathered about the students? Why?
9. I'm going to show you the prompt and your response to one of the open-ended survey items you answered last week. *[show prompt and response]* Can you take me through the response again and explain your reasoning for the aspects you thought were effective and ineffective?
10. Finally, what does it mean to you to equitably assess student learning?

Teacher Assessment Interview 2

1. Can you describe your experience so far throughout the teacher education program courses learning about assessment
2. How have your cooperating teachers assessed student learning?
3. Have your cooperating teachers explicitly discussed opinions or strategies about assessing student learning?
4. Can you describe your experiences assessing student learning in your teaching placement?

PACT Reflection – part of Teacher Assessment Interview 3

Now, I am going to ask specific questions about the focal PACT assessment – that is, the task you used to analyze student work [show or describe task].

1. Can you please take me through the structure of the assessment, what it assessed, and why you chose it.
2. Do you think that all of your students had a fair chance to show what they knew or could do on the assessment? Why or why not?
3. How did you know whether your students learned the learning objectives being assessed?
4. Do you think the assessment contributed to student learning about [the learning objective]?

Appendix B
Survey Open-ended Prompts

Assessment Plan

A. Choose one of the following science topics: Mendelian genetics, acids and bases, light and optics, or earthquakes

B. Describe in as much detail as possible how you would assess student learning during this unit and (explain why you would assess this way.

Assessment Critique

A. Read the prompt below.
B. Describe and explain to what extent they thought Ms. Sanchez's assessment practices were effective, including specific things they thought were or were not effective and describe what they would do differently
C. List other information (if any) they would like to have about the scenario to comment on her assessment practices.

Ms. Sanchez wants her 9th grade Biology students to use evidence as they explain laboratory investigation findings. She assumes that her students have never engaged in evidence-based explanations before; therefore, after a laboratory investigation on osmosis, she models different ways to use evidence, such as "I claim that…" and "the following evidence….led me to the claim…" Then, in groups of four, students orally practice explaining their findings.

Later on in the unit, Ms. Sanchez provides each student with information and data about another investigation related to osmosis. She then asks students to "write an explanation about what is happening in the investigation based on the information and data provided." Students individually complete the task and turn their response into Ms. Sanchez. She uses a rubric to score the results, which contributes to the students overall class grade. She informs each student of his/her score and keeps the actual student responses.

CONCLUSION

CONTINUING TO UNRAVEL THE COMPLEX: A PERSONAL REFLECTION

Well it's a lot more complex and tricky than it seemed back then.
-Lauren, Interview 3

Throughout this study, I have set out to unravel the complexities that exist while teachers become prepared to assess science, particularly in linguistically diverse classrooms. Furthermore, I set out to refine ways to study science classroom assessment from a teacher-centered perspective. Individually, each of the three chapters in this dissertation has contributed toward these goals, while collectively telling a cohesive story.

In the first chapter, "Unpacking the Complexity in Science Classroom Assessment: Development and Application of the Construction-Use-Equity (CUE) Assessment Expertise Framework," I conceptualized assessment expertise and translated this conceptualization into a scoring rubric to analyze teacher products. Given the dearth of literature exploring expertise in science classroom assessment, laying the conceptual foundation was an important aspect of the dissertation and can inform other studies that investigate various aspects of science classroom assessment.

The second chapter, "Rough Waters and Smooth Sailing: Charting the Changes of Secondary Science Preservice Teachers' Assessment Expertise," captured, both quantitatively and qualitatively, the ways in which the assessment expertise of secondary science preservice teachers changed during their teacher education program. The teachers encountered several key barriers that, only when successfully navigated, led to higher levels of assessment expertise. For example, the

teachers realized that they should not choose the content and form of assessment arbitrarily, but rather carefully and purposely design assessment tasks that align with what the students should be learning – i.e., the learning objectives. Although practitioner orientated publications[1] and, I would imagine, many science method courses emphasize a strong alignment between curriculum and assessment, the findings from this study provide evidence that the teachers navigated the assessment-curriculum alignment barrier. However, even though I taught the teachers about incorporating and aligning evaluative criteria (i.e., *how will you know if students have mastered the objectives?*), the teachers did not navigate this barrier as easily. Future research can explore such transitions in more depth.

In a similar vein, as reported in Chapter 3, "What about Language?: Case Studies of Preservice Teachers' Evolving Expertise in Equitable Science Assessment," Dean, Glenda, and Lauren encountered critical barriers specific to assessing science in linguistically diverse classrooms. While increasing their awareness of language's role while assessing, they experienced two primary tensions: whether to assess language use *in addition to* science content and whether they should scaffold language on assessment *as opposed to* just reducing language demands of assessment. As researchers and teacher educators move toward effective assessment-focused support for science teachers, understanding these barriers is critical. Just as science teachers benefit by knowing various alternative conceptions that students bring, and drawing on those conceptions while teaching science, teacher educators

can pay close attention to these critical barriers as they help teachers interrogate their own thoughts and resolve conflicts and struggles.

In summary, there are two key points I want to stress. First, science classroom assessment is a complex practice, one that takes considerable effort to master and that needs considerably more theoretical and methodological development. Second, while preservice science teachers can understand and enact some basic elements of this complex practice, they need more support and practice to assess science in ways that truly promote academic excellence and equity for all students. As an example, this would include assessing language use in addition to conceptual understanding and providing students with feedback on language use.

In this concluding section, I see my job as to reflect critically on what I learned – beyond the "results" – and consider my own next steps that can take the research deeper and deeper. While I began unraveling the complexities of science classroom assessment, I found some tightly wound knots that take even further unpacking. To me, this is what a successful dissertation does. The process of theory development, data collection, data analysis, and writing continuously formulates new ideas and questions that reveal new knots. Therefore, I conclude with a list of questions that I continue to think about while considering the next steps in my research agenda.

What Exactly *is* Assessment Expertise?

Being a central construct in my dissertation, the conceptualization of assessment expertise evolved throughout the study. Instead of presenting an

exhaustive list of what teacher need to know and be able to do with assessment, I attempted to extract core dimensions (i.e., Construction, Use, Equity), each of which drawing on theory and used to guide a host of teaching practices tailored to the particular student and teacher context. Of course, I had to make decisions about particular assessment practices to emphasize by considering what is most important for *classroom* assessment. For one, I left out an explicit focus on technical quality, instead deciding to emphasize the alignment of assessment to theoretical underpinnings of learning – outlined in the assessment triangle model. I also decided not to interrogate the teachers' beliefs about standardized testing or to explore how standardized testing influenced their assessment practices. Both topics are important lines of research, well documented in the literature, but not connected to how I conceptualized expertise in science classroom assessment.

I conceptualized assessment expertise as a relationship between assessment understanding and facility.[2] While I finally settled on "understanding" as meaning assessment knowledge and beliefs, I vacillated between viewing beliefs and knowledge as independent or seamlessly integrated constructs. I was also torn between whether "practices" should be viewed as a type of facility or an extension of expertise. Finally, I contemplated how "orientations" toward assessment, learning, and equity fit into expertise.

The translation of theory into instruction also concerned me. Although I organized assessment expertise into a multi-dimensional framework that I hoped could also be useful for preservice teachers, I was never reassured that the teachers

completely internalized or drew upon the CUE framework in practice – at best they took away the discrete components of the framework. Some teachers even mentioned that the "whole assessment triangle thing" confused them. Thus, I will continue refine exactly what it means to become an expertise at assessing science and translate such theory into instructional models that teachers can understand and integrate into their own practice.

What Outcomes are Important?

Even if a study does not explicitly mention an outcome variable, desirable outcomes, such as student achievement, positive attitudes toward science, and teaching practices that align with science education tenets, inform all educational research. Instantiated in the CUE rubric and coding schemes, I have implied that teachers should strive for higher levels of assessment expertise. However, I struggled with providing evidentiary support for these desired outcomes. For instance, researchers have linked the formative use of assessment to increased student learning, which warrants a desire for teachers to reach higher levels of expertise in the Use dimension. However, what about a coherent assessment system? Theoretically, aligning learning-assessment task-interpretation should yield information that is more useful for the teacher – a better representation of what students know and can do that reflects what is being taught and how. Yet, is there evidence that if teachers engage in such practices, it will eventually lead to increased science learning? More importantly, when considering assessment though an equity lens, science learning might not be all that matters. Are all students, regardless of language proficiency,

given an *opportunity* to access a rigorous science curriculum through a teacher's assessment practices? Are they developing English language proficiency in addition to learning science? As I move forward in my research, I need to ensure that I can argue for the use of the assessment practices I am studying, which means articulating exactly what outcomes are important and desirable.

How to Mix Multiple Data Sources?

A researcher can design a rubric or observation protocol, go into a classroom, and, quantitatively or qualitative, draw inferences about aspects of a teachers' assessment practices. However, such inferences would be rather limited in scope. I attempted, with varying success, to triangulate multiple sources of data to account for the variety of ways teachers demonstrate assessment expertise. For instance, in addition to audiorecording classroom observations and writing field notes, I asked teachers to reflect on their assessment practices in the final interview. I also analyzed content from their PACT commentary. However, since I was not exactly sure what I would be finding in classroom practice, I felt uneasy about having a more systematic observation protocol. Now that I have evidence of what kinds of practices teachers were implementing, an important next step of mine is to develop an observation protocol that can analyze, both quantitatively and qualitatively, assessment practices. This protocol could include checklists of teacher behaviors, evidence from interviews and other forms of reflection, and content from classroom artifacts. Furthermore, I want to refine other instruments, for example, surveys, to analyze and link to assessment practices. Finally, I struggled with how to make sense of the multiple

sources. The results of open response survey items told a different story from the interviews. How do I reconcile and explain such differences? Are they telling different (yet overlapping) pieces of the puzzle? The power of multiple methods is to draw on the strengths they each bring; however, in future research I must find useful ways to integrate those strengths.

Where does an English Learner Focus Fit Theoretical?

I struggled, and continue to struggle, with the theoretical relationship between English learners (a targeted group of students) and the Construction, Use, and Equity dimensions explored throughout the dissertation. I consistently included assessing ELs as a focal aspect of the Equity dimension. Thus, while Construction and Use really applied to assessing "all" students, that Equity dimension looked particularly at a special circumstance – what happens when the classroom is linguistically diverse. The more I analyzed interviews, surveys, teacher products, and data from classroom observation, the more I realized that all *three* dimensions are part of assessing science in linguistically diverse classrooms. Thus, ELs embody all aspects of assessment – with equity related to just one dimension. What this tells me is that I still need to do work to refine just what it means to become an expert at assessing science in linguistically diverse classrooms. I hope that more in depth studies such as the case studies presented in Chapter 3 can enrich such theory.

The Final Remark

My development through the dissertation process in ways mirrored Lauren's understanding about assessment quoted in the introduction. I evolved in my own

thinking about assessment, science learning, and equity, which complicated my initial conceptualizations and research design. I faced several obstacles as a researcher, among them, designing informative and clear research instruments, translating theory into an analytical scheme, a rubric, and instruction, as well as managing the vast amount of data. Yet, working through those obstacles brought true insight.

I have illustrated the complexity of science classroom assessment through the changes in preservice science teachers' assessment understanding and facility. I have begun to unravel the complex.

Footnotes

[1]See, for example, Wiggins, G. P., & McTighe, J. (2005). *Understanding by design. Alexandria, VA:* Association for Supervision and Curriculum Development.

[2] Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., . . . Bernbaum,D. J. (2006). Developing Expertise With Classroom Assessment in K-12 Science: Learning to Interpret Student Work. Interim Findings From a 2-Year Study. *Educational Assessment,11*(3&4), 237-263.