Summer 2017

Formative Assessment for Middle School Mathematics Instruction: An Evidence-based Approach to Evaluating Teacher Posing, Pausing, and Probing Moves

Carrie Lee Holmberg
San Jose State University

FORMATIVE ASSESSMENT FOR MIDDLE SCHOOL MATHEMATICS INSTRUCTION: AN EVIDENCE-BASED APPROACH TO EVALUATING TEACHER POSING, PAUSING, AND PROBING MOVES

A Dissertation

Presented to

The Faculty of the Educational Doctoral Program in Educational Leadership

San José State University

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

by

Carrie Holmberg

August 2017

The Designated Dissertation Committee Approves the Dissertation Titled

FORMATIVE ASSESSMENT FOR MIDDLE SCHOOL MATHEMATICS INSTRUCTION: AN EVIDENCE-BASED APPROACH TO EVALUATING TEACHER POSING, PAUSING, AND PROBING MOVES

by

Carrie Holmberg, M.A.

APPROVED FOR THE EDUCATIONAL DOCTORAL PROGRAM IN EDUCATIONAL LEADERSHIP

SAN JOSÉ STATE UNIVERSITY

August 2017

Brent Duckor, Ph.D., Chair	Department of Teacher Education
Joanne Rossi Becker, Ph.D.	Department of Mathematics and Statistics
Diana Wilmot, Ph.D.	Chief Academic Officer/Principal Yavneh Day School

ABSTRACT

FORMATIVE ASSESSMENT FOR MIDDLE SCHOOL MATHEMATICS INSTRUCTION: AN EVIDENCE-BASED APPROACH TO EVALUATING TEACHER POSING, PAUSING, AND PROBING MOVES

by Carrie Holmberg

This study involved empirical investigation of a moves-based conceptualization of teacher practices of planning, enacting, and reflecting on formative assessment (FA) in mathematics classrooms in a high-needs school district in California. A qualitative case study of six middle school mathematics teachers' practices of *posing* questions, *pausing* to foster equity of participation and quality of response, and *probing* student thinking, the study provides empirical evidence of qualitatively distinct levels of teacher posing, pausing, and probing moves. The study utilized a National Research Council-based educational assessment design framework that employed construct maps, multi-faceted items design, and scoring guides to examine teacher practice and to provide feedback protocols for teachers engaged in FA. Guided by the 2014 *Standards for Educational and Psychological Testing*, the study provides evidence for content validity and tools for future rater reliability investigations. The study found levels of teacher questioning practice, operationalized as posing, pausing, and proving moves, could be represented along generalized continua in the context of middle school mathematics instruction. The study's work toward the development of a teacher learning progression framework in the formative assessment domain has implications for establishing an empirically-based, common grammar of practice in mathematics instruction and preparation.

ACKNOWLEDGMENTS

My dissertation study would not have been possible without the critical and timely support of key personnel in the school district in which the study occurred. I would like to thank the director of educational services there who welcomed and backed my research project. I also especially thank the three principals of the schools at which the participating teachers taught for being open to my research project and for facilitating my inviting the teachers' participation.

I am grateful to the six mathematics teachers who opened up their classrooms, practices, and thinking to a stranger. The professionalism, dedication, skill, and care I witnessed them demonstrate continues to inspire me. Each teacher brought to the art and science of formative assessment an abundance of love for the young learners before them. This shone through all they did. This abundance helps make public schools places of hope, connection, and transformation for youth.

Many mentors contributed their experience and wisdom to this project. My committee members, Brent Duckor, Joanne Rossi Becker, and Diana Wilmot, individually and together embodied the best of formative assessment throughout the planning, conducting, and writing up of this study. Above all, they were open to meeting me where I was. Each is a master problem-solver. They provided much-needed feedback "just in time"—cordially and unflinchingly. I am especially thankful for their flexibility, faith, and trust.

I acknowledge and thank my chair, Brent Duckor, for his vision, commitment to my growth, perspicacity, and his fantastic ability to put concepts, principles, and practices I needed to learn into terms, chunks, and metaphors I could understand, accept, digest,

and—I like to think—do something valuable and meaningful with. Brent Duckor and I wrote a book together during my Ed.D. experience, and somehow—largely due to his capacity to shift frames quickly and completely—we made it work.

I thank the director of the Ed.D. program in Educational Leadership, Arnold Danzig, for his diligent efforts to provide the first doctoral cohort at SJSU with opportunities for intellectual and professional growth, which in my case included financial support to attend annual meetings of the American Educational Research Association in Washington, D.C. and San Antonio, Texas. This work has benefited from my experiences presenting at those conferences.

Great advice and crucial support from several people encouraged and made possible the completion of my first draft. I thank Ron and Gigi Talcott, Tom and Tina Green, Cliff and Carrie Kalinowski, Dorothy Hill-Aird, Rob Stephan, Gabrielle Moyer, Dan and Regina Sakols, and K.C. and Jennifer Chang and my mother, Sandi Kalinowski. Lois Hamor's chipper missives brought me lightness along the way. Gene Dorsa always gave me perspective and support. I thank Louise Bernbaum for her speedy feedback. My former Wilcox High School English department colleague, Misa Sugiura—long-standing friend and current water polo teammate—helped me through every challenging episode.

When life events rocked my world and affected my ability to work, Lorri Capizzi—throughout all—listened, guided, advised, prayed, accepted, and helped me to see afresh. I could not have done what I needed to do without Lorri's support. Thank you, Lorri.

My children, Hazel and Vivian, missed much time with me as I conducted and completed this work. I acknowledge their sacrifices, which were tremendous. Their

mother's becoming a more complete person through work that takes her away from them is not something they can truly understand at their ages. And yet they handled this time in our lives with grace.

My husband, my life partner, Bob Holmberg, has an incredibly big heart, nearly unfathomable stamina, and a soul-growing sense of humor. For him, this work is but one of many important endeavors we have faced together. Thank you, Bob, for your high standards, patience, and creativity. Thank you for being the one who, at critical moments on this journey, figuratively—or in actuality, as you did one long, memorable night when I was revising a document in our minivan so I could meet a deadline—handed me a flashlight. Taking risks with you expands my being.

I dedicate this work to the late Raymond C. Kalinowski, my father, who modeled hard work, clarity of thought, service to others, close reading, good humor, discipline, and grace. He told me once, "'Things' weren't always so, Care."

I am grateful to all those who helped—and continue to help—make "things" so. You are legion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

This dissertation presents the design and pilot of a performance-based assessment, known as the "P-P-P Assessment," to examine and assess teachers' posing, pausing, and probing. The study was conducted with public school middle school mathematics teachers in a high-needs district. I undertook the work from an evidence-based approach to the design of educational assessments, informed by Wilson's (2005) Constructing Measures (CM), or "building blocks", framework. In telling the story of this empirical study, and contextualizing the components of the P-P-P Assessment, I have chosen to depart from a traditional, positivist five-chapter structure. Instead, in this "design dissertation," I dedicate one chapter each—chapters 2, 3, and 4—to the three conceptual building blocks I employed to design the P-P-P Assessment, and one chapter to the presentation of the results contextualized: chapter 5, "Profiles in Practice and Feedback." Chapter 5 offers six tightly organized, targeted, and evidence-based illustrations of individual teachers' practices of a moves-based conceptualization of formative assessment, one profile for each of the teachers in my study. Focused and next steps-oriented formative feedback for each teacher follows each profile.

This first chapter is divided into six sections that serve to (a) state the problem addressed, (b) explain the motivation for the study, (c) present the research questions, (d) identify the aims of the research, (e) introduce three conceptual frameworks and one set of standards guiding the study, and (f) and describe the organization of the dissertation. Together the components of this first chapter frame the descriptions and contextualization of the empirical work that makes up the body of this dissertation.

The first section introduces the problem this work addresses: lack of well-articulated teacher learning progressions in the domain of formative assessment, and the significance thereof. To do this, this section speaks to the intersections of three topics relevant to this study: (a) teachers' practice of formative assessment as language support, (b) students' needs for academic language development, and (c) how new mathematics standards that have recently been widely adopted relate to this work. This first section also addresses the notion that teachers' practices of formative assessment influence discourse in mathematics classrooms, and the significance of this in terms of efforts to foster equity in classrooms and schools. This is critical since the P-P-P Assessment was designed with a focus on creating classrooms with more equitable spaces for learning, through helping teachers improve their practice of formative assessment.

Statement of the Problem

Extensive evidence has linked teacher practice and skill in formative assessment (FA) to increased student achievement (Hattie, 2009, 2012) in ways that have shown promise for decreasing achievement gaps between high-achieving and low-achieving students (Black & William, 1998). Educational stakeholders—including parents, community members, business leaders, policymakers, and educational leaders—have called for all students to experience high-quality instruction (Darling-Hammond, 1997) that serves to help make classrooms and schools more equitable spaces for learning. Classrooms and schools where teachers prioritize formative assessment and frequently and equitably formatively assess can productively work toward this important goal (Linquanti, 2014). Despite the recognized promise and importance of teachers engaging in skilled and

frequent formative assessment with their students, however, the ways in which teachers develop this much needed expertise and skill in enacting formative assessment practices during class is not well articulated yet. We know little about how teachers' knowledge of formative assessment and use of formative assessment practices emerge from novice to more expert "moves" in classrooms.

Unresolved issue and significance of the problem. Presently the field of education lacks well-articulated teacher learning progressions in the critical domain of knowledge and skills in formative assessment. My dissertation study addresses this gap by exploring teacher learning progression in the context of middle school mathematics instruction in three dimensions of formative assessment hypothesized: teacher-orchestrated posing of questions, pausing intentionally for a variety of reasons, and probing of student thinking. Educational assessment experts have argued that having defined learning progressions is foundational to providing feedback to learners (Alonzo, 2011; Black, Wilson, & Yao, 2011; Corcoran, Mosher & Rogat, 2009; Heritage, 2008).

In the case of teachers-as-learners, without teacher learning progressions in the critical domain of formative assessment, feedback to teachers—such as that of colleagues, administrators, or instructional coaches in professional learning communities or during structured cycles of formative observation and evaluation—cannot reach its potential. Nor can teachers' self-assessments of their practice become what they could without empirically validated learning progressions in the domain of teacher practice of formative assessment.

Once empirically validated teacher learning progressions exist in the field, they can play a foundational role in improving formative feedback to teachers. This improved formative feedback *to* teachers should help bring about improved formative assessment practices *by* teachers. As previously mentioned, research supports that skillful practice of formative assessment benefits student learning (Hattie & Timperley, 2007) and plays a role in decreasing student achievement gaps (Black & Wiliam). Together this research and these assertions form the basis of the argument that the lack of well-defined and empirically validated teacher learning progressions in the domain of formative assessment is a significant problem worthy of study.

Moreover, Heritage and Heritage (2013) have argued that the "epicenter" of formative assessment is teacher questioning, the focus of this study. By examining how six middle school mathematics teachers plan, enact, and reflect on posing, pausing, and probing moves for the purposes of instructing their linguistically and racially diverse classes of students—over 80% of whom are categorized as socioeconomically disadvantaged—and then locating the teachers' proficiency in these practices on a continuum, this empirical study adds both content and complexity to the field—in the form of pictures of individual teachers' trajectories of FA—and sharpens the discussion on finer-grained observation tools available to those interested in supporting teacher growth in the domain of formative assessment.

Good formative assessment practice as language support. Cognitive scientists, educational psychologists, experimental psycholinguists, educational researchers and practitioners have long known that language use plays a critical role in the development

of knowledge (Bodrova & Leong, 2007). Educational practitioners, as well as experts in theory and research, have also long recognized the "exquisite connections" between disciplinary subject matter content and language (Hakuta, 2017). "Good formative assessment," educational researcher and experimental psycholinguist by training Hakuta (2014) has claimed, is not only a valued and research-supported measurement tool and activity, it is also a critical instructional practice to support students' language development. According to Hakuta (2013) and other experts, the practice of good formative assessment is language support for students.

Because educators recognize the way language affects content knowledge, they have turned to formative assessment practices—and "formative assessment technologies"—as a way to better understand students' readiness to engage with the language of the content being taught. Technological advances have included apps and tools marketed to teachers specifically "for giving formative assessments" (Molnar, 2017). Teachers have appropriated many of these new technologies to support their students' content language development.

Though technology can play an important role in helping to develop students' content language and disciplinary thinking, good formative assessment practice as language support need not incorporate it. In today's culturally, linguistically, and economically diverse classrooms, it is arguably more critical to support students by using an instructional framework for formative assessment that helps teachers plan and improvise instruction that highlights, uncovers, encourages, and sharpens their own and students' use of academic language (Duckor, Holmberg, & Rossi Becker, 2017). This study

provides additional empirical evidence of an instructional framework that does just that: the FA moves framework (Duckor, 2014), in the context of middle school mathematics classrooms.

The role of language in the standards of mathematical practice. The relatively recent Common Core State Standards in Mathematics (CCSSM), which were adopted by the California State Board of Education in 2010 and modified in 2013, also recognize students' comprehension and use of content language as critical for developing knowledge in a discipline. Concomitant with the CCSSM are eight Standards of Mathematical Practice (CCSSI, 2010). Among other activities, these "math practices" call for all K-12 students engaging in mathematics to reason and explain (math practices 2 and 3) as they "make sense of problems and persevere in solving them" (math practice 1). Table 1 lists the eight standards of mathematical practice adopted with the CCSSM by the California State Board of Education in 2010 (CDE, 2013, 2014).

Table 1*: Standards of Mathematical Practice*

Standard	Practice
1	Make sense of problems and persevere in solving them.
2	Reason abstractly and quantitatively.
3	Construct viable arguments and critique the reasoning of others.
4	Model with mathematics.
5	Use appropriate tools strategically.
6	Attend to precision.
7	Look for and make use of structure.
8	Look for and express regularity in repeated reasoning.

Educational stakeholders are motivated to support teachers in instantiating these math practices in classrooms. This is done for many reasons, including reasons related to new methods of testing and systems of accountability being implemented in California.

Annual testing of K-12 public school students since spring 2014 has been aligned to the CCSSM (Smarter Balanced, 2017). These new assessments "portend significant change" (p. 184), argues mathematics education scholar, Alan Schoenfeld (2015). Yet the question for teacher educators, researchers, and educational stakeholders on "how to prepare students and teachers for such change" remains (Schoenfeld, p. 184).

Students' need for support in developing academic language. All students need opportunities to develop specialized academic language that is associated with learning in content domains (O'Hara, Zwiers, & Pritchard, 2012). For students who are learning English, however, producing and developing academic language during school particularly supports learning and academic success. English learners need significant opportunities to engage in structured academic talk in classrooms (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006). Students' academic language development has been cited as a major contributor to gaps in achievement between English learners and their English-proficient peers (Anstrom et al., 2010; Francis et al., 2006).

Teachers' practices of formative assessment have a role to play in addressing students' academic language development needs. Scholars have recognized the potential of the inherently dialogic process of formative assessment between teacher and students to provide linguistic-minority/high-poverty students in particular with multiple opportunities to develop academic language while engaging in meaningful, discipline-specific practices (Abedi, 2010; Linquanti, 2014; Ruiz-Primo, Solano-Flores, & Li, 2014). Hakuta (2013). Other experts (Moschovitch, 2013, 2015; Spanos, Rhodes, Dale, & Crandall, 1988; Zwiers, O'Hara, & Pritchard, 2014) have pointed to the need for teachers

to attend closely to students' academic language production as an integral aspect of supporting students' mathematical reasoning processes.

No studies to date, however, have documented teacher learning processes and the effects of specific observation frameworks and protocols on teacher learning—particularly in relation to students' academic language production and middle school mathematics classes. While this is not the primary focus of the present study, the conceptualization of formative assessment employed in the study, the FA moves framework (Duckor, 2014), "lends itself to sustaining a focus on the development of academic language for all students, which is critical to fostering equity in mathematics learning and teaching" (Duckor, Holmberg, Rossi Becker, p. 336). The present study represents foundational work that would be needed to support thoughtful attempts to document the effects of specific teacher observation frameworks and protocols on teacher learning in relation to students' academic language production and development.

Formative assessment, classroom discourse, and equity. Classroom discourse patterns are affected by teachers' practices of formative assessment. Further, classroom discourse affects students' confidence in mathematics and their sense of themselves, as equals, in the classroom. Darragh (2013) found that when students' contributions are valued in class discourse—as good formative assessment practice can—the confidence students expressed in their mathematical competence increased. Boaler (2002) found that teaching and learning practices of secondary teachers can significantly reduce linguistic, ethnic, and class inequalities in their schools. This would include, one could infer, in-class formative assessment practices that influence classroom discourse.

More recently, Boaler and Sengupta-Irving (2016) found that middle school students' engagement, achievement, and enjoyment of mathematics increased with equity-focused teaching. Duckor and Holmberg (2017) have argued that "good formative assessment practice" not only supports students' language development and helps teachers in their instructional decision making, it also seeks to achieve results in equity of opportunity and engagement.

Skillful practice of formative assessment practice should positively influence learning outcomes and students' social-psychological perceptions and functioning. Given the importance of middle school mathematics' achievement to their mathematics course-taking trajectories and performance in subsequent years (Balfanz, 2009; Finkelstein, Fong, Tiffany-Morales, Shields, & Hoang, 2012), pursuing knowledge about how teachers can better practice formative assessment in the context of middle school mathematics instruction is an endeavor of potential significance. Gaining, documenting, and communicating such knowledge through the collection and analysis of relevant empirical evidence is one of the purposes of the present study.

Motivation for the Study

Extensive research supports this study's investigation into teacher questioning, with a focus on posing, pausing, and probing. Teachers' practices of in-class formative assessment have well-documented advantages to student learning (Black & Wiliam, 1998; Hattie & Timperley, 2007; Hattie, 2009, 2012; Ruiz-Primo & Furtak, 2006; van Zee , Iwasyk, Kurose, Simpson, & Wild, 2001). The central goal for teachers in this process is to construct questions and make decisions in real time that support and further

student learning. Though recognized as at the heart of good teaching (Hattie, 2012; Hattie & Timperley, 2007; Heritage & Heritage, 2013), formative assessment remains understudied and under-theorized (Allal & Pelgrims, Ducrey, 2000; Duckor, 2014; Erickson, 2007).

In addition, research on formative assessment that engages with teachers' practices, particularly in relation to questioning during class, has also paid little attention to culturally and linguistically diverse populations (Abedi, 2010; Duckor, 2014; Jiang, 2014; Ruiz-Primo, Solano-Flores, and Li, 2014). This is despite the facts that teacher questioning is considered by experts to be a high-leverage practice (Ball, Sleep, Boerst & Bass, 2009) and that public school classrooms continue to grow more culturally and linguistically diverse (Klein, 2015).

Focusing on teachers' questioning practices—especially within multilingual classrooms—is also significant because researchers have established that beginning and experienced teachers' practice of questioning can be positively influenced through various interventions (Black, Harrison, Lee, Marshall, & Wiliam, 2004; Jacobs, Lamb, Philipp, & Schappelle, 2011; Ong, Lim & Ghazali, 2010). Yet how teachers develop expertise and skill in posing, pausing, and probing during class is still largely anecdotal and impressionistic. While researchers have established, within mathematics instruction at least, that attending to student thinking is a prerequisite for deciding how to respond to students' verbal contributions in class (Jacobs, Lamb, Philipp, & Schappelle, 2011; Moyer & Milewicz, 2002; Sahin & Kulm, 2008), little is known about how teachers

develop questioning skills that ultimately influence student thinking in the classroom (Duckor, 2014; Erickson, 2007; Moyer & Milewicz, 2002).

All these arguments—arguments that support investigation into teachers' posing, pausing, and probing in linguistically diverse classrooms and the construction and validation of methods that can validly and reliably evaluate teachers' practices—motivate this study. The goal of this study is to provide further empirical evidence that helps teachers develop expertise in these three dimensions of formative assessment practice, including by generating individualized formative feedback for participants.

Research Questions

The research questions of this study aim to aid the design of a performance-based assessment for teachers, the P-P-P Assessment, which attempts to reliably and validly evaluate their practices of posing questions, pausing for a variety of purposes, and probing student thinking in the context of middle school mathematics classrooms. The order of the questions in this study reflect the phases of development of the P-P-P Assessment. The study poses three research questions (RQs):

RQ1. Is there a continuum of practice from novice to more expert knowledge and skills in teachers' posing questions, pausing for a variety of purposes, and probing student thinking in middle school mathematics? Can we assess teachers' knowledge and skills on this continuum? This question addresses the extent to which the three proposed dimensions of formative assessment appear in the research literature. Further, it examines the extent to which there might be patterns between this study's findings and empirically validated descriptions of teacher practice of these dimensions.

RQ2. How can levels of practice be represented along a generalized continuum in mathematics teaching and learning at the classroom level? This question explores how construct maps can articulate variation among and between teachers and responses to items that attempt to elicit evidence regarding the proposed dimensions of formative assessment practice. The complexity of evaluating what happens between the extremes of "novice" and " expert" practice is the primary focus of the investigation in this study. In addressing this question, choices regarding how practice can be represented are explicated. For example, what are the ramifications of the decision to describe each dimension in terms of three facets—planning, enactment, and reflection—as a way to represent practice?

RQ3. Can teacher practices of posing, pausing, and probing along any or all of these proposed dimensions of practice be reliably and validly evaluated? This question seeks to examine the accumulation of validity evidence from several sources with attention to the quality of the inferences that can be made when the P-P-P Assessment is used as intended.

Aims of the Research

The aims of research questions 1 through 3 are to assist in the design of the P-P-P Assessment, to inform possible future iterations of the assessment, and to provide evidence of its functioning. In principled educational assessment and evidence-based design of assessments, reliability and validity are central concerns of any instrument used to draw inferences about an underlying latent construct; they must be taken into careful consideration at the start and throughout any assessment design project. Taken all

together, the research questions of this study will help to ascertain the extent to which the P-P-P assessment functions well enough to merit further study.

The knowledge and understanding gained from the study is also expected to have potential utility for those who engage with preservice and inservice teachers to support their professional development. Such stakeholders include preservice teacher educators, teacher induction specialists, instructional coaches, principals, other teacher evaluators, and professional development providers to inservice teachers.

The study yields information valuable to these stakeholders who wish to formatively assess teachers and to support teachers' growth and development of practices that show promise in helping realize more equitable instruction during class. The data and findings of the study may also serve a role in future arguments for supporting further efforts to explore the multi-dimensionality of teacher practices of in-class formative assessment from a moves-based perspective.

This study is also important because the study adds empirical data and analysis on a particular assessment approach to teacher practice. This assessment approach can be characterized as taking a formative, cycle-of-inquiry approach as opposed to a summative, exam-event approach to teacher practice. In the present context, the "Era of Accountability," where methods of assessing teacher performance in comparison to professional standards are of interest to educators, parents, policy makers, community members and business leaders; introducing, using, and reflecting on this study's particular assessment approach may make potentially valuable contributions to the dialogue on evaluating and supporting teachers. Assessment approaches that foreground

formative purposes, as this study's approach to assessment does, are presently not the norm. Such approaches are, however, needed. These aspects of the study merit attention.

Another aspect of this study that is significant is that the methods employed and the perspectives from which learning progressions are developed matter (Shavelson, Moss, Wilson, Duckor, Baron, and Wilmot, 2010). Conducting the work toward developing a *teacher learning progression* from two specific stances—the evidence-centered assessment design stance and the moves-based framing of formative assessment stance—will make it more likely that the work's contribution to the field will be innovative, grounded in evidence, aligned with first principles of assessment, reflect up-to-date theories of cognition and teacher learning, and hold potential for positively influencing classroom discourse amongst teachers and students in classrooms.

Frameworks and Standards Informing the Study

Three conceptual frameworks and one set of standards guided the work: (a) the National Research Council's "assessment triangle" (Pellegrino, Chudowsky, & Glaser, 2001), (b) Wilson's Constructing Measures—or "building blocks" —approach to assessment design, (c) Duckor and Holmberg's (2017) moves-based conceptualization of formative assessment, and (d) the 2014 testing standards published jointly by the American Education Research Association, the American Psychological Association, and the National Council on Measurement in Education. This section introduces them and outlines their role in the work.

The National Research Council's "assessment triangle". The first conceptual framework that guided the work of this study is known as "the NRC assessment triangle,"

a graphical representation and mental model used to communicate a panel of experts' thinking about constructing and evaluating measures. Of the three conceptual frameworks I outline in this section, I introduce this one first because it is the most foundational.

In 2001, experts in the field of educational measurement, assessment, and testing on the Committee on the Foundations of Assessment, sponsored by the National Research Council (NRC), published a report marking an important step in research into the cognitive foundations of knowledge domains and how they relate to educational assessment. The experts identified the key concepts that any study of these knowledge domains must contain and represented the concepts and their inter-relationships with a mental and visual model called the "assessment triangle" (Pellegrino, Chudowsky, & Glaser, 2001), shown in Figure 1. The assessment triangle represents the foundational elements of an assessment and their relation to one another through its three vertices of *cognition*, *observation*, and *interpretation*.

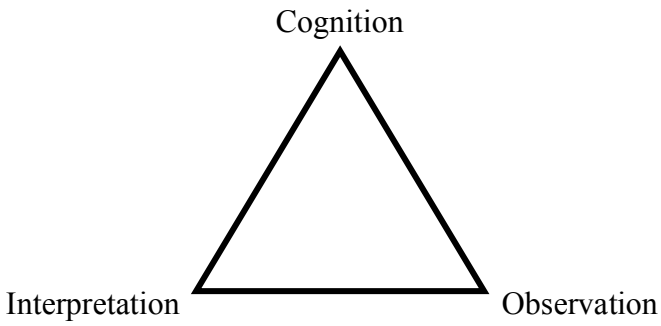

Figure 1: Assessment triangle (adapted from NRC).

The three elements, represented by the vertices of the assessment triangle—*cognition, observation, and interpretation*—should function in synchrony to support (a) the purpose of the assessment and (b) arguments for the validity and reliability of the inferences that can be drawn from it when it is used as intended (Pellegrino, Chudowsky, & Glaser, 2001). The first vertex of the assessment triangle, *cognition*, refers to the theory of cognition or learning in a domain under study. The term *construct*, used to describe the underlying human trait being measured, is represented by this vertex.

The second vertex of the assessment triangle, *observation*, describes the set of prompts, tasks, or situations expected to elicit demonstrations of the construct or underlying human trait under study. Assessment experts frequently refer to the contents of this vertex as "the items." Experts assert that the tasks or items that human subjects are prompted to respond to in an assessment are deliberately and carefully chosen (Pellegrino, Chudowsky, & Glaser, 2001). They are chosen to function well in generating responses from subjects from which sound inferences regarding the construct or constructs under investigation can be made. Experts have many methods for doing so and use clear and distinct rules for ensuring connections between the *cognition* and *observation* vertices.

The third vertex of the assessment, *interpretation*, represents the framework behind examining the evidence collected from the *cognition* and *observation* vertices in light of the constructs from the *cognition* vertex in principled ways. The NRC (2001) committee sees the *interpretation* vertex as encompassing "all the methods and tools used to reason

from fallible observations" (p. 48). This vertex is commonly referred to as "score interpretation" (AERA, APA, NCME, 2014).

A crucial point about the three vertices of the assessment triangle, and relevant to the work behind this dissertation, is that each vertex, or element, of the assessment "must make sense on its own *and* connect to each of the other elements in a *meaningful way* to lead to an effective assessment and sound inference [italics added]. [All] three vertices of the triangle must work in synchrony (NRC, 2001, p. 49)." Also pertinent to the topic of this dissertation: when developing an assessment, the NRC predicts, it will "almost certainly be necessary for [assessment] developers to go around the assessment triangle several times, looking for mismatches and refining the elements to achieve consistency" (NRC, 2001, p .51). Assessment design is an iterative process, informed by data, analysis, and expertise.

Wilson's Constructing Measures approach to assessment design. The second of three conceptual frameworks that guided the work of this study is Wilson's "building blocks" approach to designing assessments, also known as the Constructing Measures (CM) framework. I explain this conceptual framework second since the "blocks" in the "building blocks" approach are proxies for the vertices of the NRC's assessment triangle.

The CM framework is an item-response modeling approach to constructing measures in the tradition of evidence-based designs of assessments. When using the building blocks approach during assessment design or measurement construction, "measurers" employ qualitative and quantitative investigations to inform their decisions regarding design choices. Measurers interrogate the soundness of the logic chains that link evidence

with arguments supporting the reliability and validity of inferences drawn from the measure. This dissertation will use the terms “assessment designer” or “assessment developers” in place of Wilson’s references to “measurers.”

Four building blocks make up the Constructing Measures framework. The blocks are: (a) construct maps, (b) items design, (c) outcome space, and (d) measurement model. The present study employed the first three. See Figure 2 for a visual representation of the building blocks, which should work in an integrated fashion to support the quality of the instrument, or assessment, being designed. Establishing sound arguments regarding reliability and validity, based on evidence, are central to the quality of the assessment, which is why in the visual representation of the building blocks in Figure 2 “reliability evidence” and “validity evidence” reside in the center and double arrows from each building block point to the them, together.

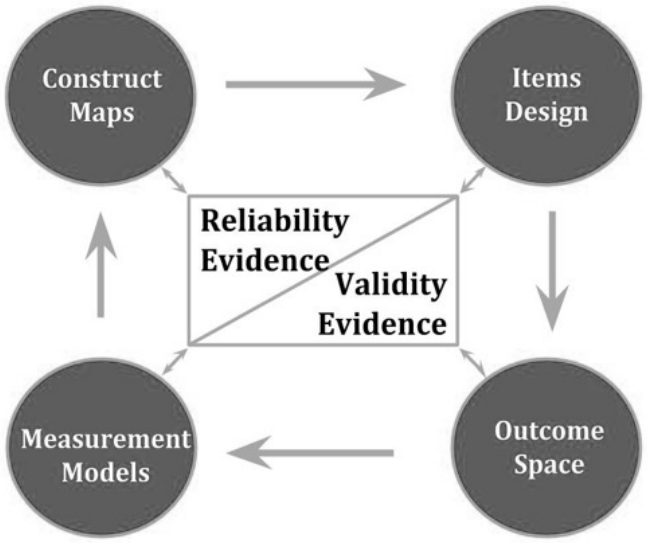

Figure 2. Schematic representation of relations among "building blocks" (adapted from Duckor, Draney, & Wilson, 2009).

Wilson's building blocks framework is analogous to the "elements" framework that was presented in the NRC report and is represented by the NRC's assessment triangle. Just as the NRC represents its framework with the assessment triangle's vertices for *cognition*, *observation*, and *interpretation*, Wilson's Constructing Measures framework is represented by the four building blocks. Each building block corresponds to an "element" or vertex of the NRC assessment triangle.

While the NRC assessment triangle has three vertices, there are four building blocks. Wilson uses two blocks for the interpretation vertex, to represent significantly different aspects of the interpretation vertex, the differences of which are critical in employing the

item response modeling approach and to manifesting the affordances this approach lends to the work of assessing and the process of developing quality assessments.

Wilson's framework shares the NRC (2001) report's emphasis on the need for all the elements—the four "building blocks"—to work in synchrony. Finally, as the NRC (2001) report emphasizes the need for concepts of "consistency," "meaning," and "refining," in the design, performance, and evaluation of assessments, Wilson's framework emphasizes the need for "quality control" concepts, particularly reliability and validity. The four building blocks should work together to support the reliability and validity of the inferences that can be drawn from the assessment when used as intended.

Wilson's approach specifies beginning to build a measure by creating "construct maps" to represent and articulate the theory of *cognition* regarding the construct being assessed. This is the first building block: construct maps. Next, to make *observations*, assessment developers carefully and deliberately create and/or select "items" that aim to elicit evidence of performance of the construct. This is called *items design,* the second building block in Wilson's approach. The third building block is known as the *outcome space*. It corresponds with the *interpretation* vertex and encompasses all that goes into deliberate and strategic defining of the system that describes how assessment takers' responses to items will be interpreted and scored. Expert assessment designers use their procedural knowledge of how to categorize observations and to score them as indicators of the construct under investigation in ways that support the assessment to function as intended.

Wilson's framework conceptually "splits" the assessment triangle's single *interpretation* vertex into two: the *outcome space* and *measurement model* building blocks. *Measurement model*, the fourth building block, in simplest terms, is concerned with checking how item responses "match" back to the construct under investigation. This entails relating the scored data back to the construct map. This building block requires quantitative analysis of test takers' responses in strategic ways to examine the relationship between the degree of the construct possessed by respondents and responses to items.

Experts choose which interpretative—and always quantitative—strategy or strategies they will use according to context and problems to be solved. Their measurement model choices can enable the affordances of item response modeling to be manifested. This includes being able to use patterns of item responses to describe—predict—expected variation in levels of student understanding of, for example, mathematical functions (Wilmot, 2008).

The choice to use Wilson's approach holds significance. Using Wilson's Constructing Measures framework, and employing the first three of four building blocks to inform the work of this study increases the likelihood that the development of the P-P-P Assessment aligns with recognized principles of assessment design such as those expressed in the 2001 NRC report and those in the 2014 *Standards for Educational and Psychological Testing* (AERA, APA, NCME).

A moves-based conceptualization of formative assessment. The third of three conceptual frameworks that guided the work of the present study and that I introduce and

outline in this section is the "FA moves" framework (Duckor, 2014). The FA moves framework conceptualizes formative assessment as a dynamic, pedagogical process of moves between teacher and students. Using this conceptual framework helps teachers to learn more about students' understandings and to productively respond to those understandings—not merely "misconceptions" or "wrong" answers—during instruction. Using the FA moves framework requires acts of planning, instructing, and reflecting on soft data to make better decisions. The FA moves framework inherently places a premium on feedback loops in classroom talk, building up of repertoires of auditory and verbal skills, and providing instructional space for students to use academic language and register during lessons.

The framework identifies seven moves accessible to novices and useful to more expert teachers across content areas (Duckor & Holmberg, 2017). These FA moves include: *priming, posing, pausing, probing, bouncing, tagging,* and *binning*. See Figure 3. Each move ties together instruction and assessment practices by requiring evidence of student engagement and visible routines for eliciting thinking. At the core of the FA moves framework are deep, sustained routines related to questioning: posing questions, pausing for think time, and probing on initial student responses to invite elaboration.

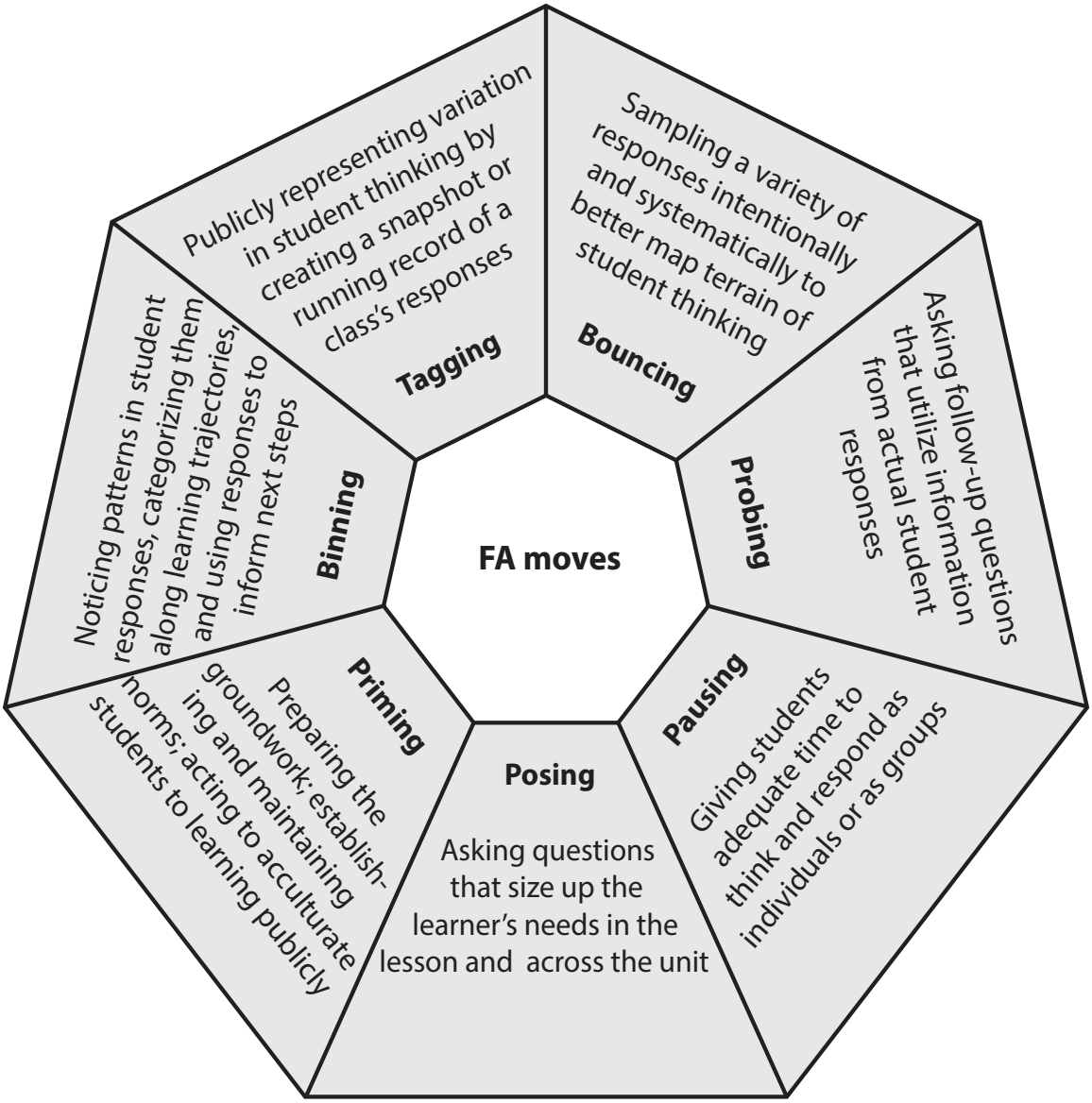

Figure 3. FA moves wheel (Duckor, Holmberg, & Rossi Becker, 2017). Reprinted with permission from *Mathematics Teaching in the Middle School,* copyright 2017, by the National Council of Teachers of Mathematics. All rights reserved.

This moves-based framing of formative assessment practice is rooted in other theoretical frameworks. It links to Sadler's (1989) focus on feedback loops in classroom discourse and builds on Wiliam's (2007) framework of "key strategies" of assessment for

learning to provide "'window[s] into [student] thinking'" (p. 1069). In mathematics, it aligns with Ball and associates' (2009) concept of high-leverage practices which enhance opportunities for mathematical reasoning and discussion. Now that the three conceptual frameworks informing the present study have been introduced and outlined, I foreground the set of standards most critical to deciding the orientation I used while planning, collecting, and analyzing evidence about the design, initial development, and preliminary functioning of the P-P-P Assessment.

The 2014 testing standards. The most up-to-date standards on assessment in education are the 2014 Standards for Educational and Psychological Testing, sponsored by the leading educational research, psychological research, and measurement organizations for professionals in their fields: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The Standards provide criteria for the development and evaluation of tests and testing practices. Their purpose is to provide guidelines for assessing the validity of interpretations of test scores for the intended uses (p. 1).

The *Standards* provide a frame of reference to ensure that relevant issues concerning educational and psychological testing are addressed. The *Standards* not only codify the most informed and researched-based assertions of experts in the field regarding what "bars" should be met in educational and psychological testing, the *Standards* declare:

> All professional test developers, sponsors, publishers, and users should make reasonable efforts to satisfy and follow the *Standards* and should encourage others to do so. All applicable standards should be met by all tests and in all

> test uses unless a sound professional reason is available to show that a standard is not relevant or technically feasible in a particular case (p. 1).

The concepts of validity and reliability/precision are foundational to testing and assessment. Reflecting the significance of validity, reliability/precision, and fairness in testing, the *Standards* places explanations of these concepts and the standards related to them first in Part I, "Foundations."

Validity. According to the Standards, validity is a unitary concept, "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (p. 14). Rather than distinct types of validity, the Standards refer to "types of validity evidence" [italics added]. This distinction is critical, with implications for experts and practitioners committed to assessment as a process of evidentiary reasoning. Types of validity evidence need to be examined in careful relation to one another.

The Standards recognize five types of validity evidence: (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence for validity and consequences of testing. This study included work related to the first two of the five types of validity evidence, though the evidence based on response processes was limited to the exit survey, a few questions in the reflection session, and three items in the pre-lesson enactment survey. Cognitive labs and "think alouds," standard protocols for collecting evidence based on response processes, were not conducted as part of the present study.

Task analysis, one method commonly used to augment the collection of validity evidence, was conducted on the performance tasks of P-P-P Assessment. The results of the task analysis of two items that anchored collection of planning- and reflecting-related evidence: (a) the common lesson planning template and (b) the video-stimulated recall (VSR) protocol, are presented in chapter 3, Items Design.

Procedures to gather validity evidence based on test content and to support the argument for validity were conducted, but details regarding the expert content panel review, the analyses of related teacher performance assessments, and the alignment checks with relevant professional standards that occurred are not explicated or reported in this draft of the dissertation. Exclusion of the latter three types of validity evidence—evidence based on internal structure, evidence based on relations to other variables, and evidence for validity and consequences of testing—was deemed appropriate in this study given the scope of the project, its employment of only the first three of Wilson's building blocks, and the assessment's potential and foreseeable uses in the near term.

Reliability/precision. The 2014 Standards use the term "reliability/precision" to denote a general notion of consistency of scores across instances of a testing procedure. The Standards use the term "reliability/precision" over "reliability" to avoid ambiguity of meaning, since in the measurement literature the term "reliability" has been used in different ways (p. 33).

Reliability/precision is always important in measurement (AERA, APA, NCME, 2014). Yet the need for precision increases as the consequences of decisions based on the

measurement grow in significance. Moreover, reliability/precision has implications for validity. The *Standards* state:

> Reliability/precision of data ultimately bears on the generalizability or dependability of the scores and/or the consistency of classifications of individuals derived from the scores. To the extent that scores are not consistent across replications of the testing procedure (i.e., to the extent that they reflect random errors of measurement), their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited (p. 34-35).

The *Standards* outline principles behind and procedures for evaluating reliability/precision, for examining factors that affect it, and documenting and reporting evidence for reliability (such as appropriate standard errors, reliability or generalizability coefficients, or test information functions). According to the *Standards* evaluating reliability/precision should involve, where appropriate, examinations of inter-rater, test-retest, alternate forms, and internal consistency reliability/precision. Although RQ3 addressed the notion of reliability and the efforts in designing the P-P-P Assessment strove for reliability/precision, and though generalized scoring guides aligned with each of the construct maps were employed with an intra-rater agreement protocol, the present study does not report findings related to reliability/precision.

Organization of the Dissertation

The remainder of this design dissertation is organized into five chapters. Rather than present a review of the literature in a stand-alone chapter, reviews of the relevant literature are incorporated into chapters 2, 3, and 4. Each of these middle chapters corresponds to one of the three building blocks employed in the study in the design of the

P-P-P Assessment. Chapter 2 is dedicated to construct maps, chapter 3 to items design, and chapter 4 to outcome space.

Each of these three middle chapters follows a similar structure. First, each chapter begins by defining the concept that is represented by the building block to which that chapter is dedicated. Next, each chapter contextualizes the application of that concept/building block to the P-P-P Assessment within the field of K-12 education and educational research. This middle section of chapters 2, 3, and 4 answers the question, "How are the constructs maps/items design/outcome space of the P-P-P Assessment positioned in the field?" Finally, the third and last section of each these chapters describes the results of applying the concepts and principles of that building block to the design challenges of the P-P-P Assessment and, where relevant, presents key design decisions and revisions made during the study.

More specifically, chapter 2, the construct maps chapter, highlights how qualities of construct maps make utilizing them a good choice for the immediate and longer-range purposes for designing the P-P-P Assessment. It introduces the idea that revising the construct maps employed in the process of building and piloting an assessment is normal, necessary, and can serve to strengthen the evidence-based argument that score interpretations from an assessment can make. Chapter two describes the drafting and revision processes of the three construct maps for the P-P-P Assessment and presents the final version of the posing construct map at study's end (see appendix A for all three).

Chapter 3, the items design chapter, outlines characteristics of good performance tasks, explains why they are used in evaluating the practices of teachers, and presents the

items design of the P-P-P Assessment. Chapter 3 features (a) a timeline of the items of the P-P-P Assessment in the sequence in which they were presented to subjects, (b) the results of the task analysis conducted on two key items of the P-P-P Assessment: a planning-focused item and a reflecting-focused item, and (c) the table of specifications for the P-P-P Assessment.

Specifically chapter 4, "Outcome Space," explains the qualities of a sound and useful outcome space and three approaches commonly used to develop one. The chapter examines the corresponding concept of an outcome space in a set of tools developed by the Educational Testing Service for a committee of the Council of Chief State School Officers interested in supporting teachers to practice and improve their practices of formative assessment: Wylie and Lyon's conceptualization of "Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice" for K-12 teachers, known as FARROP. The final section of chapter 4 explains early and final versions of the scoring guides created for the P-P-P Assessment since they are central to the design of its outcome space (see appendix B for all three final versions of the scoring guides).

Chapter 5, "Profiles in Practice and Feedback," presents concise and evidence-based descriptions of the posing, pausing, and probing practices and planning, enacting, and reflecting skills of the six teachers in the study. Data derived from participants' engagement with the P-P-P Assessment were used to locate their practices along the generalized continua articulated in this study, thereby addressing aspects of each the three research questions asked in this study. Individual formative feedback that identifies areas

for growth related to each teacher's planning for, enactment of, and reflection on posing, pausing, and probing follows each evidence-based description of practice. The individualized formative feedback is tied to the responses generated from each teacher's engagement with the P-P-P Assessment and is intended to fall within each respondent's zone of proximal development in the practice of formative assessment as conceptualized in this study.

The dissertation closes with chapter 6, which discusses the limitations, future directions, and implications of the study. Limitations derive from sample size and selection bias, use of video, and curriculum and content effects. Chapter 6 recommends specific directions for future work that will continue to establish the argument for evaluating teachers' formative assessment practices by means of the P-P-P Assessment in order to support its wider use and adoption.

Chapter 2: Construct Maps

The three sections of this chapter together (a) define the concept of the construct map, (b) situate where and how the three construct maps employed in the P-P-P Assessment are positioned in the field of K-12 education, and (c) and present the content of the P-P-P Assessment's construct maps by narrating their developmental process and explaining key decisions I made while drafting and revising them. This chapter's first section explains what construct maps are and identifies the measurement tradition from which they come. It highlights the qualities of construct maps that make utilizing them a good choice for the immediate and longer-range purposes of designing the P-P-P Assessment. It introduces the idea that revising the construct maps employed in the process of building and piloting an assessment is both normal, and necessary, and can serve to strengthen the evidence-based argument that score interpretations from an assessment can make. An explanation of the drafting of the posing, pausing, and probing construct maps and the ways in which this supported the items design of the P-P-P Assessment is explored in chapter three.

Construct Maps: A Definition

Mark Wilson (2005) developed the concept of the construct map within a measurement and assessment framework. A construct map—sometimes called a *progress variable*—is a visual representation of the variable, or construct, an assessment designer is aiming to assess. A construct map is meant to help conceptualize how assessments can be designed to relate to theories of cognition (Wilson, 2009, p. 2). Experts in educational assessment (Pellegrino, Chudowsky, & Glaser, 2001) and the 2014 *Standards for*

Educational and Psychological Testing (AERA, APA, NCME) call for grounding educational assessments in up-to-date, research-supported and sound theories of cognition. Taking a developmental perspective on learning and the assessment of learning is one way to begin to do so.

A construct map embodies a developmental perspective on assessment of student achievement and growth (Wilson, 2009). In simplest terms, a construct map is a "well thought out and researched ordering of qualitatively different levels of performance focusing on one characteristic" (Wilson, 2009, p. 3). In this study, "teacher pausing" during lessons is one such characteristic. The other two characteristics—or hypothesized dimensions of formative assessment practice—are teacher posing and probing. Therefore I drafted and revised three construct maps for this study and the design of the P-P-P Assessment.

Affordances of the nature of the complexity of construct maps. According to Wilson, "a construct map is intended to be a somewhat less complex concept than a learning progression" (2009, p. 2). At the same time, a construct map's own complexity, and the nature of this complexity, imbues construct maps with qualities that make them (a) essential to the design of sound assessments, (b) useful for instructional decision making (when instructional practices are linked to vetted and relevant construct maps), and, potentially, (c) helpful in efforts aiming to integrate instruction and accountability (Wilson, 2009).

Recognizing these latter two aspects of construct maps helps to contextualize the work of this study and to highlight its potential significance. Further, sound assessments

are designed in particular contexts for particular purposes (Pellegrino, Chudowsky, & Glaser, 2001; Wilson, 2005). These latter two aspects of construct maps ("b" and "c") relate to two purposes of the P-P-P Assessment. The primary of these two purposes for the P-P-P Assessment (and its construct maps) is to be useful in linking instructional practices for developing teachers, whether they are pre-service teachers or in-service teachers. Another purpose of the P-P-P Assessment—after further development and validation—could include playing a role in efforts that aim to integrate instruction of teachers and accountability. The nature of the complexity of construct maps is an essential component of the capacity of the P-P-P Assessment to be able to achieve both of these purposes, given requisite conditions that would allow the P-P-P Assessment, when used as intended, to play a role in doing so.

Significance of the potential cross-curricula interpretability of construct maps. The nature of the complexity of construct maps I am referring to is that construct maps are at once general enough to be interpretable within a particular curriculum—and possibly across curricula, according to Wilson (2009) —and specific enough to guide an assessment designer as she develops the items design and outcome space (or scoring guides) of an assessment. To the extent that construct maps demonstrate capability—and success—in being interpretable across curricula, they hold possibility for facilitating the creation of methods, Wilson has argued, "for large-scale assessments to be linked in a principled way to what students are learning in classrooms, while at least having the potential to remain independent of the content of a specific curriculum" (2009, p. 3). As the designer of the P-P-P Assessment, I am interested in this possibility since the

"students" I have developed the P-P-P Assessment and its posing, pausing, and probing construct maps for are teachers, and, further, they are teachers who are not engaged in learning from a "specific curriculum."

Future studies might apply the posing, pausing, and probing construct maps—and some future iteration of the P-P-P Assessment—to groups of pre-service teachers exposed to a "specific curriculum" in common. This might be pre-service teachers enrolled in the same single subject or multiple subject credential program. Other future studies might employ the construct maps with groups of inservice teachers exposed to the "same" "specific curriculum" in person or online. These might be teachers participating in a professional development program as a cohort. But this condition—that the "students" for whom the construct maps are developed have in common that they are engaging with a "specific curriculum"—is not present in this study.

This condition should be kept in mind, along with the purposes of the P-P-P Assessment, as I present the content, features, and characteristics of the posing, pausing, and probing construct maps and their iterations in this chapter. The value of the utility of posing, pausing, and probing construct maps, it seems to me, resides in their capability to be usefully and productively applied to teachers who do not necessarily share engagement with a specific curriculum. Indeed, nearly everything I have done creating and iterating the posing, pausing, and probing construct maps has sought to enhance their utility and performance for application across a wide variety of teachers and teaching contexts, even as this study has only applied them to the context of teachers teaching middle school mathematics lessons. I have not lost sight of the larger—and future—goal

of validating these construct maps with a larger, broader, and more diverse teaching population.

Significance of creating, using, and iterating construct maps, an item response modeling approach to assessment. The structure and function of construct maps come from an item response modeling approach to assessment (Lord, 1980, Rasch, 1960, Wright, 1977; Wright & Masters, 1981). Item response modeling (IRM) was developed to enable comparisons among test takers who take different tests and to enable comparisons among test items whose parameters are estimated using different groups of test takers (Pellegrino, Chudowsky, & Glaser, 2001). A powerful affordance of IRM is that it is possible to predict the properties of an assessment from the properties of the items of which it is composed.

Being able to "talk" about what tends to happen with specific assessment items allows for meaningful score interpretations to result from assessment situations in which different teachers respond to different items. This becomes increasingly important to be able to do—and do reliably—when large numbers of teachers are assessed, and when the teachers being assessed vary widely in terms of their proficiency regarding the latent construct. The latent construct, in this case, is a teacher's practice of formative assessment as defined by the instructional "moves" of posing, pausing, and probing.

Choosing an approach to designing the P-P-P Assessment that (a) embodies a developmental perspective on assessment of student proficiency and growth—or *teacher* proficiency and growth, since the P-P-P Assessment is designed to assess teachers, not students—and (b) possesses the capability to (potentially) be taken to scale was important to me. I wanted to base any assessment of teachers' instructional practice I was designing

on deep empirical work that delved into both the broader nature and salient particulars of teachers' development, largely to distinguish the P-P-P Assessment from other assessment and evaluation tools pertaining to teachers' instructional practice prevalent in K-12 education. The available evidence on many of these other frameworks, tools, rubrics, and observation protocols does not indicate how they are based on deep empirical work or strong theories of human cognitive development.

Features of a construct map. The two most important features of a construct map, according to Wilson (2005), are that a construct map (a) communicates a coherent and substantive definition of the content of a single dimension of the variable being assessed, and (b) that it is composed of an underlying continuum that describes how an individual shows "more" or "less" of this variable. This is accomplished by ordering respondents (individuals who take the assessment) and their responses to items from greater to less—or more skillful to less skillful—and qualitatively grouping them into an ordered series of levels (Wilson, 2005, p. 26).

A construct map has two sides. Each side speaks to a different aspect of the variable, or construct, being assessed. There is a *respondents* side and a *responses to items* side. These are separated on a construct map since it is critical, when employing this particular item response modeling approach to the designing of an assessment, for an assessment designer to move beyond the *p-prim* that "the latent construct *is* the items." (Duckor, Draney, & Wilson, 2009; Wilson, 2005). (*P-prims* are intuitive ideas based in everyday experience and not expert conceptions of a topic.)

This necessary separation of *respondents* from *responses to items* on a construct

map—from the very first stage of designing an assessment—can assist an assessment designer in developing a more sophisticated understanding of assessment. This, in turn, can assist an assessment designer in designing an assessment that will function better to support the quality of the inferences that can be drawn from it. It is also critical for an assessment designer to move beyond simply—and incorrectly—equating the latent construct with the responses to items, as some novice assessment designers will do (Duckor, 2005; Duckor, Draney, & Wilson, 2009; Wilson, 2005). Again, the structure of a construct map, and the function it plays in Wilson's "building blocks" approach to the design of assessments, works to challenge such a misconception and to keep it from degrading the quality of an assessment as it is being designed and developed.

Descriptions of *respondents* and *responses to items* are also separated on a construct map in order to facilitate being able to describe and locate respondents in terms of their possession of the "amount" of the latent construct independently from the descriptions of the qualitatively ordering of responses to items. This separation is critical to being able to—in a later phase of an assessment's development—characterize and mathematically describe the relationship between the degree of construct possessed by respondents and item responses and analyze how the items are functioning to distinguish respondents. This is the most important purpose of using an item response modeling approach for the design of an educational assessment over other approaches that could guide the design of an educational assessment.

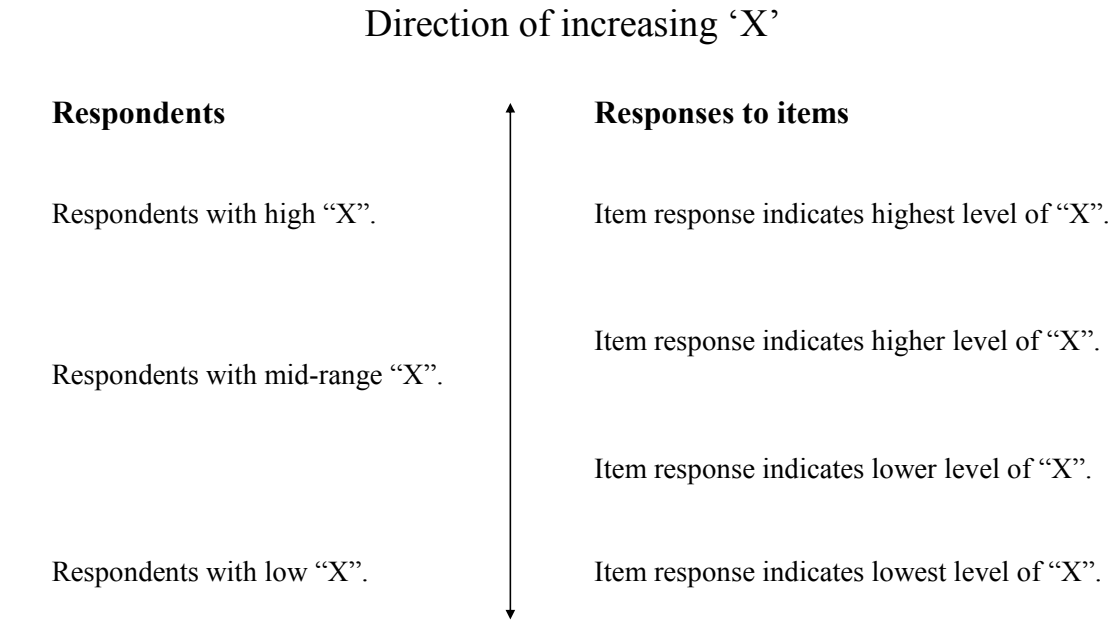

Figure 4. Generic construct map (adapted from Wilson, 2005).

Both sides of a construct map are ordered similarly from “high” to “low”—or from “more” to “less”—from top to bottom for the variable being assessed. By convention, the left-hand side of a construct map is the *respondents* side. The left-hand side indicates qualitatively distinct groups of respondents, ranging from those with high amounts of the variable to those with low amounts. By convention, the right-hand side is the *responses to items* side. This side indicates qualitative differences in responses to items (or tasks), ranging from responses that indicate high amounts of the variable to those that indicate low amounts (Wilson, 2005).

Role of construct maps during phases of design of an assessment. Construct maps play a critical role at every stage or phase of the design of an assessment. Assessment designers create and use construct maps at the initial stage of designing an assessment in order to focus on the essential feature of what they are attempting to assess: in what way does an individual show "more" or "less" of the construct, or variable, being assessed. As the process of designing and developing the assessment continues, designers may revise the construct maps they are using as their analysis of incoming empirical evidence impels them to do so. This was the case for me in this study. I will present and explain the revisions in this chapter.

Skilled assessment designers will try to align their items design and outcome space (scoring guides) with the construct maps (Duckor, Draney, & Wilson, 2009). During the development of their assessment, they will employ their construct maps in ways that serve to strengthen—or perhaps intentionally weaken—inferential links between specific aspects of the measurement framework (Duckor, Draney, & Wilson, 2009). This is done to bolster the reliability and validity of the inferences that can be drawn from the assessment when it is used as intended.

Number of construct maps an assessment employs. The design and development of an assessment may employ one or several construct maps. Since a construct map "can be said to be a unidimensional latent variable" (Wilson, p. 7), and some constructs are more complex than this, it may be necessary to represent each dimension of the construct one is attempting to measure one at a time in order to see each as a construct map. Wilson recommends (2005, p. 7) this strategy of taking each one individually, and calls doing so

part of the process of "variable clarification" (p. 38). Clarifying the variable one is attempting to assess is a requirement of designing and building an assessment that will work well.

I have employed this strategy with the design of the P-P-P Assessment since I hypothesize teacher practice of formative assessment as a construct encompassing several dimensions, which can be described in terms of the instructional moves *priming*, *posing*, *pausing*, *probing*, *bouncing*, *tagging*, and *binning*. In this study, I have chosen to focus on the dimensions of formative assessment defined by the posing, pausing, and probing construct maps.

Where and How Posing, Pausing, and Probing Construct Maps Fit in the Field

As previously mentioned, construct maps are not learning progressions, nor are they rubrics. Others in the field of K-12 education who have articulated claims about teacher questioning—posing and probing—or claims about how teachers provide and support students with "wait" or "think" time—pausing—have not employed construct maps in their schemes.

Most of the most well-known schemes in the field that describe teacher practice and include some aspects of the dimensions of formative assessment hypothesized in this study—posing, pausing, and probing—have used rubrics featuring three or four levels to do so. These are

- Danielson's Framework for Teaching (FFT; 1996, 2007, 2011, 2013),
- the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State

School Officers' (CCSSO) "Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice" (FARROP) (Wylie & Lyon, 2013),

- the Interstate New Teachers Assessment and Support Consortium's (INTASC) Learning Progressions for Teachers (LPfTs) 1.0 (CCSSO, 2013), and
- the Teaching for Robust Understanding (TRU) Framework (Schoenfeld, 2014).

The INTASC's scheme for organizing claims about teacher practice is an exception. It refers to the organization of its document as "learning progressions for teachers," not rubrics. The LPfTs 1.0 document includes "shifts" teachers are expected to make as they progress from one level to the next in each three-level "progression" articulated. Traditionally, rubrics do not include explicit descriptions of shifts expected to occur in knowledge and skills between levels of performance, though they may suggest this implicitly. Including these expected shifts differentiates the INTASC scheme from the others.

All four schemes apply generally to teachers of all subject areas. Schoenfeld's TRU Framework originally was intended to apply only to mathematics instruction. The framework has been adapted, however, to apply to teaching of all subject areas and learning environments of all kinds—classrooms, schools, and organizations (Schoenfeld, 2015).

Table 2 identifies the focus of each of these four schemes and briefly describes the items design of each. 1 also characterizes the extent and quality of any validation work done regarding the framework, as well as its relationship to posing, pausing, and probing

(P-P-P) as defined in this study.

Table 2: *Focus, Items Design, Validation, and Degree of Relation to the P-P-P of Four Prevalent Schemes in K-12 Education that Articulate Claims About Teacher FA Practice*

Framework	Focus of Constructs	Items Design	Validation	Relation to P-P-P
FFT	K-12 general	22 four-level rubrics	extensive	high
FARROP	FA	10 four-level rubrics	none	medium
LPfTs 1.0	K-12 general	9 progressions: 3 levels, 2 shifts each	none	low
TRU	Originally math, also K-12 general	5 three-level rubrics	new	medium

Note. FFT = Framework for Teaching (Danielson, 2013); FARROP = Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice (Wylie & Lyon, 2013); LPfTs 1.0 = Learning Progressions for Teachers 1.0 (CCSSO, 2013); TRU = Teaching for Robust Understanding (2014); P-P-P = P-P-P Assessment.

Although Dylan Wiliam's matrix of five "key strategies" for formative assessment (2007) is widely known in K-12 education, this chapter does not address Wiliam's matrix since it does not explicitly articulate claims about teacher questioning or teachers providing and supporting students with "wait" or "think" time.

This next section describes the relation of each framework, set of tools, or scheme to the posing, pausing, and probing construct maps for this study. It begins with Danielson's Framework for Teaching, followed by the Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice and then INTASC's Learning Progressions for Teachers 1.0. Finally, this is followed by Schoenfeld's TRU (math) scheme and its rubrics.

Framework for Teaching. The Danielson Framework for Teaching is the most widely used instructional framework in the U.S. and has the greatest amount of validation research associated with it. Since its introduction in 1996, Danielson's Framework for Teaching has "by merit and by default become part of the foundation for efforts to improve teacher evaluation in the U.S. (Milanowski, 2011, p. 3)." Danielson's Framework for Teaching (FFT) saw increased engagement with school districts after the U.S. Department of Education announced its competitive grant program, Race to the Top (RTTT), in July 2009. States' abilities to win RTTT funding was linked to their having certain educational policies in effect, specifically teacher performance evaluation systems that were based on multiple measures of educator effectiveness (USDE, 2009).

Danielson's FFT is also significant since it was used in the Bill and Melinda Gates Foundation-funded Measures of Effective Teaching (MET) Project, which involved over 3,000 educators in six urban school districts across the U.S. from 2009-2012. Of the several observation tools used in the Project—which included the Protocol for Language Arts Teacher Observations (PLATO), the Mathematical Quality of Instruction (MQI), and the UTeach Teacher Observation Protocol (UTOP) (Kane & Staiger 2012)—only the FFT and the Classroom Assessment Scoring System (CLASS; LaParo & Pianta, 2003; Pianta, Hamre & LaParo, 2007) were used in the Project to rate the videotaped lessons of both mathematics and English Language Arts (ELA) lessons across all of the grade levels included in the project (Kane, McCaffrey, Miller, & Staiger, 2013). Further, the FFT was the only framework used in the analysis of MET data done by Kane, McCaffrey, Miller, and Staiger (2013), a study that concluded the FFT had "predictive validity" (Danielson

Group, 2016).

In "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment" Kane et al (2013) concluded that:

> a composite measure of effectiveness (with appropriate controls for prior student achievement) can identify teachers who produce higher achievement among their students. Moreover, the actual impacts on student achievement (within a school, grade, and subject) were approximately equal on average to what the existing measures of effectiveness had predicted. These are causal impacts, estimated with random assignment. (p. 38)

In this validation study, 19.9% of the students in the evaluated classrooms were designated English language learners by their school districts (Kane et al., 2013, p. 17). This contrasts with 20-46% designated English learners attending the three schools in my study.

The Framework for Teaching (FFT), in the words of its author, is "a definition of teaching quality and a classroom observation system designed to enrich deliberations in school systems on ways of improving instruction" (Ferguson & Danielson, 2014, p. 99). This definition of teaching quality is represented as "dimensions" of teaching performance which are accompanied by a set of rubrics. Though the term "dimension" is used in regard to the FFT (e.g., MET literature ["Danielson's Framework for Teaching for Classroom Observations"] describes Domain 3 of the FFT, Instruction, as "measur[ing] several *dimensions* of instructional quality including communication, discussion techniques, ability to engage students, use of assessment during instruction, and flexibility and responsiveness" [italics added] (MET, p. 2), this is a casual use of the term. To my knowledge, no evidence of analyses done to confirm or dispute the accuracy of these "dimensions" is available. From a measurement perspective, use of the term

"dimension" implies that analyses, such as factor analysis or item response theory-based analyses, have been conducted to gather data that is used to advance the validity argument for an assessment, i.e., gather validity evidence based on internal structure. Without presentation of this validity evidence, use of the term "dimension" in regard to the FFT is colloquial in nature, rather a fact-based use that is reflective of prevailing standards in the field of assessment and educational measurement.

Each rubric in the FFT describes four levels of performance. The FFT is intended to provide a common language for conversations about teaching practice and is meant to apply to all subject areas, pre-kindergarten through grade 12. In 2007, the Danielson Group published specialized rubrics for educators working in specialist positions, such as school counselors and librarians.

The FFT divides teaching into four domains: (a) planning and preparation, (b) the classroom environment, (c) instruction, and (d) professional responsibilities. Each domain has specific performance "components" and "elements" nested within them. There are 22 components across the four domains and 2-5 elements per component. See Table 3.

Each component in the framework includes an associated list of "indicators". Each component has a rubric or rating scale that describes the four performance levels in terms of observable teacher or student behavior. The four performance levels are (a) unsatisfactory, (b) basic, (c) proficient, and (d) distinguished. Each rubric also articulates "critical attributes" and "possible examples" to illustrate what practice can look like in a range of settings at each of the performance levels. Bolded text in Table 3 indicates the

components of the FFT that express a noteworthy degree of intersection with posing, pausing, and probing.

Table 3: *Domains and Components of the Framework for Teaching*

Planning and Preparation	Classroom Environment	Instruction	Professional Responsibilities
Components 1a-f	Components 2a-e	Components 3a-e	Components 4a-f
1a Demonstrating knowledge of content and pedagogy	2a Creating an environment of respect and rapport	3a Communicating with students	**4a Reflecting on teaching**
1b Demonstrating Knowledge of students	2b Establishing a culture for learning	**3b Using questioning and discussion techniques**	4b Maintaining accurate records
1c Setting Instructional outcomes	**2c Managing classroom procedures**	3c Engaging students in learning	4c Communicating with families
1d Demonstrating knowledge of resources	2d Managing student behavior	**3d Using assessment in instruction**	4d Participating in a professional community
1e Designing coherent instruction	2e Organizing physical space	**3e Demonstrating flexibility and responsiveness**	4e Growing and developing professionally
1f Designing student assessments			4f Showing professionalism

Note. **Bolded text** denotes area of most connection between FFT and posing, pausing, and probing.

Validation of the FFT. The Danielson Group claims (2016) that "Each component of the Framework for Teaching has been validated by the Measures of Effective Teaching (MET) study" and that "The Framework for Teaching has been found to have predictive

validity." This claim should be considered in light of the 2014 Testing Standards that assert that validation is an ongoing process (AERA, APA, NCME, 2014, p. 19) and that the validity of test score interpretations depends not only of the uses the test scores, but specifically on the claims that underlie the theory of action for these uses (pp. 19-20). According to the Standards, then, school districts or other bodies using the FFT should keep in mind alongside claims of "predictive validity" both the intended use of the FFT—that is, "to enrich deliberations in school systems of ways of improving instruction" (Ferguson & Danielson, 2014, p. 99) and that when a test user proposes an interpretation or use of test scores that differs from those supported by the test developer, the responsibility for providing validity evidence in support of that interpretation is the responsibility of the user (p. 13). "Predictive validity" concerning the FFT's use in the MET study does not necessarily imply transportable, applicable, or corresponding validity—predictive or otherwise—concerning the FFT's use in other school districts.

Incorporation of teacher planning and reflection. The posing, pausing, and probing (P-P-P) construct maps feature claims about teachers' planning, enactment, and reflection in relation to the move at each of the five levels articulated in the maps. They do so systematically, in a balanced way. This embedded integration of teacher planning and reflection in the construct maps contrasts markedly with the FFT's representation and incorporation of teacher planning and reflection.

Planning. In the FFT, teacher planning is addressed only in its Domain 1: Planning. A noteworthy exception in the FFT, one relevant to the purposes here of comparing the FFT with the P-P-P Assessment's construct maps, is the FFT's assertion that a "true mark of

[teacher expertise] is skill in eliciting evidence of student understanding" and that "this is not a hit-or-miss effort, but is planned carefully in advance" (p. 72). This appears in the definition of the element "Monitoring of student learning", one of four elements that comprise Component 3d, "Using Assessment in Instruction," within Domain 4, Instruction.

Reflection. The FFT remarks on teacher reflection in its Component 4a, "Reflecting on Teaching," which is one of five components within its Domain 4, Professional Responsibilities—with one exception. The only other references to teacher reflection on practice in the FFT outside of Component 4a, "reflecting on teaching," appear in the rubric for Component 3e, "demonstrating flexibility and responsiveness" (within Domain 3, Instruction) in the "Critical Attributes" section of the rubric. See Table 4.

Table 4: *References to Teacher Reflection in the Rubric for Component 3e, "Demonstrating flexibility and responsiveness," of the FFT*

Performance Level	Relevant "Critical Attribute" among 4-5 Attributes per Level
Unsatisfactory—1	In reflecting on practice, the teacher does not indicate that it is important to reach all students.
Basic—2	In reflecting on practice, the teacher indicates the desire to reach all students but does not suggest strategies for doing so.
Proficient—3	In reflecting on practice, the teacher cites multiple approaches undertaken to reach students having difficulty.
Distinguished—4	In reflecting on practice, the teacher can cite others in the school and beyond whom he has contacted for assistance in reaching some students.

Note. (adapted from FFT, 2013, pp. 78-79).

Extent of alignment. Content of the P-P-P construct maps aligns with FFT Domains 1-3 which are respectively: (1) Planning and Preparation, (2) Classroom Environment,

and (3) Instruction. Each of the 15 components of FFT Domains 1-3 is addressed in the P-P-P construct maps, with one exception: presently no explicit references to the organization of physical space of classroom (FFT Component 2e of Domain 2, Classroom Environment) are made in any of the P-P-P construct maps.

Alignment is not correspondence. This considerable alignment between the FFT and the P-P-P construct maps does not, however, imply a one-to-one correspondence between the two, even within areas that align. The P-P-P construct maps articulate claims regarding teacher practice in the areas of planning, enactment, and reflection in relation to each FA move that are not addressed in the FFT. The P-P-P construct maps articulate more detail about teacher planning and reflection concerning posing, pausing, and probing (or their proxies in the FFT) than the FFT does—with one exception. The role of student questions during lessons is expressed more prominently in the FFT than it is in the P-P-P construct maps.

Relationship between posing and FFT. The FA Moves Framework, through its posing construct map, defines posing as questioning and focuses on teachers' planning, enactment, and reflection on the formulation, deployment, and uses of teacher questions in class. To a lesser degree, posing also includes the extent to which teachers plan for, use, and reflect on student questions.

Posing figures prominently in the FFT. One of the framework's 22 components is dedicated largely to the uses of posing in instruction. Component 3b, "Using questioning and discussion techniques" is the second of five components that comprise the FFT's domain on Instruction. Another indication of the priority of posing for the FFT appears in

the role "questioning" plays as one of the two instructional strategies focused on in the framework (p. 59). As the FFT notes in its opening sentence of its description of Component 3b, "Using questioning and discussion techniques" (p. 59): this decision to focus on questioning as a central strategy reflects their [questioning and discussion techniques'] central importance to teachers' practice."

Similar to the FA Moves Framework, the FFT values questioning as a technique to "deepen student understanding" and "promote student thinking"…"rather than serve as recitation, or a verbal 'quiz'" (p. 59). Both frameworks recognize the importance of questions in encouraging students to make connections. Both frameworks further recognize the value of teachers posing questions for which they do not know the answers. Of the three FA Moves, *posing*, *pausing*, and *probing*, the strongest correspondences between the two frameworks appear around *posing*. As highlighted previously, the FFT has more frequent and explicit references to student questions and student questioning than the posing construct map has.

Relationship between pausing and FFT. As defined in the pausing construct map, pausing refers to "wait time" or "think time" to assist engagement and cognitive processing for students and the teacher too. Pausing need not necessarily imply silence in the classroom as a whole, or silence regarding a student or group of students, although pausing frequently features "quiet" moments intended to encourage reflection.

Similar to the pausing construct map, references to pausing—or proxies for pausing—in the FFT are tied to teachers' purposes. The FFT explicitly names teacher pausing in connection to checking for student understanding (p. 28), supporting students' vocabulary

development (p. 57), and making teachers' "high quality" question posing more effective (p. 60). From only four total references to *pausing*—or wait/think time—in the FFT, it can still be soundly concluded that *pausing*'s main role, as articulated in the FFT, is to "provide students with sufficient time to think about their responses, to reflect on the comments of their classmates, and to deepen their understanding" after teachers have asked a question of "high quality" (p. 60). This explanation of the role of *pausing* appears in the FFT's definition of the first of three elements that make up Component 3b, "Using questioning and discussion techniques" in Domain 3, Instruction. Its appearance there reflects the FFT's recognition of *pausing*'s importance in supporting student cognition, reflection, and meaning-making.

Both the pausing construct map and the FFT argue that the mark of "proficient" instruction ("proficient" is language from the FFT, not the pausing construct map) lies in the teacher's "effective use of wait time" (FFT, 2013, p. 63). The pausing construct map delves more deeply into teachers' reasons for *pausing* and the characteristics of *pausing* in class. For example, characteristics of expert *pausing*, pausing at the "extended abstract" or highest level articulated on the construct map, include highly and intentionally differentiated pauses orchestrated during class.

Relationship between probing and FFT. According to the probing construct map, probing is a move enacted by teachers and students alike, whose main purpose is to encourage elaboration of student thinking in order to make that thinking visible to teacher and students. Frequently, probing is instantiated in classrooms when teachers are seen asking follow up questions to individual students during whole class instruction, small

group time, or independent student work time. The probing construct map also includes the ways in which probing is supported by teachers, including actions teachers take to support student-to-student (S-2-S) probing.

This next section addresses the following three aspects of probing during instruction: (a) student-to-student probing, (b) the amount of student elaboration or explanation during class, and (c) teacher "follow-up" actions to student responses/performances elicited by probing.

Student-to-student probing. Both the FA Moves Framework and the FFT agree that when students challenge one another's thinking during class, this indicates advanced teacher practice. The probing construct map identifies this as student-to-student probing and incorporates such observations of classroom interactions into its descriptions of higher levels of teacher practice regarding probing. The FFT, without naming this particular example or similar descriptions of behavior as "probing," nonetheless includes the concept of student-to-student probing among the "critical attributes" evidence in the highest level of teacher practice regarding questioning and discussion techniques. In "Distinguished" performances of teacher practice (Level 4 of the rubric for Component 3b) in the FFT, "Students invite comments from their classmates during a discussion and challenge one another's thinking" (p.63).

Amount of student elaboration or explanation during class. An important purpose of probing expressed in the probing construct map is to encourage students to elaborate on their thinking, to explain their thoughts further. The probing construct map recognizes many potential benefits associated with student elaboration and explanation. This

includes the potential for better instructional decision making by the teacher and advancement of student learning: both the probee's learning and the learning of other students witnessing, listening, and responding to the probee's elaboration.

The probing construct map describes how teachers might use evidence/data gleaned from student elaboration and explanation. At the highest level of the probing construct map, "extended abstract," teachers would "progressively use" what probing elicits "to advance student responses toward the learning target." Although one might argue that the FFT addresses potential uses of data indirectly in the FFT's Component 3e, "Demonstrating flexibility and responsiveness", it remains that the FFT addresses student explanation explicitly in one area: Component 3c, "Engaging Students in Learning" of Domain 3, "Instruction."

The references regarding student explanation in this area reflect the importance in the FFT of instruction that "requires" students to explain their thinking. In the FFT, "Unsatisfactory" teacher performance does not invite or require students to explain their thinking." The FFT asserts that "Basic" teacher performance indicates "few…materials and resources…ask students to explain their thinking." whereas "Proficient" teacher performance demonstrates that, "students are invited to explain their thinking as part of completing tasks." "Distinguished" teacher performance provides evidence that "lesson activities require high-level student thinking and explanations of their thinking" (pp. 68-69).

Teacher "follow-up" actions to student responses/performance. Probing involves teachers following up on students' initial responses to questions posed. The FFT

addresses this aspect of probing in three "possible examples" in Component 3b, "Using Questioning and Discussion Techniques (pp. 62-63), providing illustrations of: a teacher who "doesn't follow up" on a student's wrong answer ("Unsatisfactory" teacher performance); a teacher who "does not follow up when [a] student falters" while explaining his reasoning "for why 13 is a prime number"; and a teacher who asks students to find textual support for their answer to the question 'Why do you think Huck Finn did…?' and "to explain their thinking to a neighbor" (pp. 62-63).

The FA Moves Framework, too, addresses probing as "following up." The probing construct map, however, also delves into how teachers approach probing and how they might seek to incorporate student responses they anticipate as well as responses they do not anticipate. The probing construct map also delves deeper than the FFT does into teachers' (stated/reported) intentions and capacities to uncover misconceptions while probing. Finally, the probing construct map articulates much more detail about teachers' plans and in-class actions to make use of what has been elicited, uncovered, and made visible via probing than the FFT does about what has been elicited by "follow up questions", the main proxy for probing in the FFT.

Formative Assessment Rubrics, Reflection and Observation Protocols (FARROP). This section reviews the set of rubrics, reflection and observation protocols known as FARROP and how they relate to posing, pausing, and probing as articulated in the construct maps for this study and the P-P-P Assessment. FARROP is an abbreviation for "Formative Assessment Rubrics, Reflection and Observation Protocols."

History and purposes. A subgroup of the Council of Chief State School Officers, the Formative Assessment for Teachers and Students (FAST) State Collaborative on

Assessment and Student Standards (SCASS), or FAST-SCASS, commissioned the Educational Testing Service (ETS) to create a teacher peer observation protocol for K-12 teachers that focused on formative assessment practices. The resulting set of rubrics and observation and reflection tools, referred to as FARROP in this document, were authored by Caroline Wylie and Christine Lyon.

The committee members of the 10-state FAST-SCASS then provided feedback on and suggestions for FARROP, which was then reviewed by teachers in FAST-SCASS and published in 2013. From the perspective of the Constructing Measures framework, using the system to formatively evaluate teachers' practices of formative assessment is likely to be instantiating a *p-prim* about assessing a construct embedded in the work: that the rubric is the construct. This is simply not the case. Avoiding this problem is one of the advantages of developing tools to evaluate and support teachers' practices from an item response modeling approach such as Wilson's Constructing Measures, or "building blocks," approach to designing, developing, and refining assessments and assessment suites, especially ones intended to be used for purposes of formative evaluation.

Wiley and Lyon were careful to assert that the observation tools and rubrics were not developed for summative evaluation. Further, they have asserted that the tools should not be used for summative evaluation until: (a) studies of "their validity and reliability" have been conducted, (b) a training and certification system for observers and has been created, and (c) a process for monitoring observer accuracy "on an ongoing basis" has been developed (p. 4).

Characteristics of and relationship to posing, pausing, and probing. In FARROP, ten rubrics represent ten "dimensions" of formative assessment practice. Though this system uses the word "dimension," its use does not signify that validity evidence based on examinations of the assessment's internal structure is available. No evidence that factor or item response theory-based analyses have been conducted, for example, to confirm or dispute these "dimensions" is available. From a measurement perspective, without presentation of the validity evidence supporting the internal structure of the rubrics' ten dimensions, the use of the word "dimension" is premature and misleading.

Each rubric describes both the teacher role and the student role in a particular formative assessment "dimension." In each rubric, there are four levels that describe implementation of the practice from "incomplete" or novice to expert: beginning, developing, progressing, and extending. The rubrics are intended to avoid judgment at the level of expertise and indicate, instead, the level of implementation of a particular aspect of practice.

Each level provides two to five sentence-long descriptions. For example, the "beginning level" of the rubric for "dimension five," "Feedback Loops During Questioning," includes two descriptors: "The teacher asks none or very few questions designed to encourage classroom discourse during the lesson," OR "The teacher asks questions from students, but discourse focuses on a statement of correct or incorrect rather than deeper/meaningful exploration of ideas." The "extending" level of this rubric has three descriptors:

- The teacher asks questions designed to encourage classroom discourse consistently throughout the lesson and integrates questioning and

discussion seamlessly into instruction.

- The teacher and students consistently build on other students' responses, clarify student comments, push for more elaborate answers, or engage more students in thinking about the problem.

- Extended feedback loops are used to support students' elaboration and to have students contribute to extended conversations. Classroom discourse is characterized by the consistent use of feedback/probes that encourage deeper/more meaningful exploration of ideas.

The ten "dimensions", i.e., —observation rubrics—of FA practice in FARROP can be clustered into groups.

The organizing principle behind the groups reflects a belief shared by educational experts for decades: namely, that teachers need to help students answer three questions: (1) Where am I headed? (2) Where am I now? (3) How do I close the gap? These questions provide a rich conceptual framework for formative assessment.

Ramprasad (1983) first articulated these three questions regarding feedback and Wiliam (2004, 2005) explicitly tied the questions to five key strategies in formative assessment. Eight of the dimensions/rubrics are aimed at these three questions: two engage with "Where are we headed?", three engage with "Where are we now?" and three engage with "How to close the gap?"

The "dimensions"/observation rubrics are also organized to reflect the way that formative assessment practice occurs for a *purpose*—using evidence to adjust instruction—and that it occurs *within* a supportive learning context. Figure 5 depicts FARROP's ten "dimensions"—or aspects—of formative assessment practice according to the three questions each addresses, for a certain purpose and within a particular context.

FA practice occurs within a supportive learning context.

IX. Collaboration

(1) Where are we headed?	(2) Where are we now?	(3) How to close the gap?
I. Learning Goals	**III.** Tasks & Activities that Elicit Evidence of Student Learning	**V.** Feedback Loops During Questioning
II. Criteria for Success	**IV.** Questioning Strategies that Elicit Evidence of Student Learning	**VI.** Descriptive Feedback
	VIII. Self-Assessment	**VI.** Descriptive Feedback
⇩	⇩	⇩

The purpose of FA practice is to adjust instruction.

X. Using Evidence to Inform Instruction

Figure 5. FARROP conceptual framework organizing ten "dimensions" of FA.

In addition to FARROP's relationship to Ramprasad and Wiliam's work, two of its ten "dimensions" intersect with posing, pausing, and probing: "dimensions" V and IV. The former looks into feedback loops during questioning. The latter explores teachers' questioning strategies for eliciting evidence of student learning and considers four areas of teacher questioning as a way to elicit evidence of student learning: (a) the frequency,

range, and timing of teacher questions; (b) the use of wait time; (c) the questioning strategies that reach students and the number of students reached; and (d) the extent to which the teacher's questioning strategies—including probing—help the teacher to make inferences about student progress.

"Dimension" IV also shares aspects with the probing construct map (e.g., that at higher levels teachers make productive use of student responses), and with aspects of pausing and probing (e.g., that at lower levels teachers provide inadequate wait time; and that at higher levels teachers probe for more information as necessary, to make inferences about student progress and to adjust/continue instruction accordingly).

INTASC's Learning Progressions for Teachers (LPfTs) 1.0. A set of "teacher learning progressions" intended for general application to all K-12 teachers was published in 2013 by the Interstate New Teacher Assessment and Support Consortium (INTASC), a consortium organized and supported by the Council of Chief State School Officers (CCSSO). The authors of these "learning progressions" view them as developmental progressions, valuable for supporting teacher development, not teacher evaluation (CCSSO, 2013).

Despite this characterization, however, this set of descriptions of teacher practice is a set of rubrics. They are not learning progressions in the sense using the moniker "learning progressions" implies. Learning progressions need to be developed by methods recognized in the field of assessment. By definition, construct maps, which articulate an underlying theory of growth or development in the latent traits or constructs under investigation, always are used to develop learning progressions. Learning progressions

need empirical validation to be determined such, preferably through longitudinal studies of learners.

These "learning progressions for teachers" are meant to work in conjunction with the INTASC Model Core Teaching Standards (CCSSO, 2011) to help improve teachers' ability to teach to college- and career-ready standards. The "progressions," like the model core teaching standards to which they are aligned, outline principles and foundations of teaching practice that cut across all subject areas and grade levels.

There are nine "progressions." Each of the ten INTASC Model Core Teaching Standards has an associated "progression," with the exception that Standard 1: Learner Development and Standard 2: Learning Differences together are represented by only one progression: "Learner Development and Learning Differences."

Each "progression" describes the increasing complexity and sophistication of teaching practices across three levels, which are identified simply by the numbers, 1, 2, 3. The authors of the "progression" chose not to name the levels of development in order to "avoid confining teaching practice to a 'box' that labeled performance" (p. 14).

Each "progression" includes three to four indicators that point to teaching standards of performances, knowledge, and dispositions that are all likely to be observed during instruction and instantiated in teaching practice. Each "progression" also articulates two "shifts" per indicator (there are two to four indicators per progression) that succinctly describe (in 10 to 20 words) the overall qualitative knowledge and skills necessary for teachers' practice to move from a level 1 to a level 2 and from a level 2 to a level 3. To help teachers move from one level to the next, the "teacher learning progression" also

provides illustrative examples of professional learning experiences below each "shift."

Relationship between LPfTs 1.0 and posing, pausing, and probing dimensions hypothesized. Two of the INTASC "teacher learning progressions" include content that intersects, to a limited degree, with one of the hypothesized dimensions of teacher formative assessment the P-P-P Assessment aims to evaluate: posing. These are Standard 6: Assessment and Standard 8: Instructional Strategies. Wait time, or pausing, is not explicitly referenced in the INTASC Learning Progressions for Teachers 1.0 (CCSSO, 2013) document at all. Probing is only referenced in the document in a section that lists indicators of teacher performance of Standard 8: Instructional Strategies. This is a reprise of the standards, not a part of the "teacher learning progression."

Posing is explicitly referenced in the Progression for Standard 8: Instructional Strategies in the descriptions of levels 1 and 2:

> The teacher poses questions that elicit learner thinking about the information and concepts in the content areas as well as learning application of critical thinking skills such as inference making, comparing, and contrasting (level 1).
>
> The teacher models higher order questioning skills related to content areas (e.g. generating hypotheses, taking multiple perspectives, using metacognitive processes) and engages learning in activities that develop these skills (level 2) (pp. 39-40).

The only other explicit reference to teacher posing in the INTASC "progressions" document that corresponds to teacher posing as articulated in the posing construct map is a mention of a question log. This appears in the portion of the document that offers illustrative examples of experiences that might help teachers move from one level to the next. The authors suggest that teachers "maintain a log of questions used in teaching to self-assess the variety, relevance, and rigor of questioning strategies" (p. 40). This

example corresponds, approximately, to an important part of this study: data collected on teacher planning was examined for evidence of teachers' use of questioning schemes such as question banks or question maps as a way to assist them with posing and probing. Data collected on teacher reflection was examined for awareness of, reflection on the use of, and/or commitment to try such tools. This study hypothesized that teachers who referenced question maps, banks, shells, or schemes explicitly during planning, enactment, or reflection would score higher in their enactment of posing and probing than teachers who did not.

The "progression" for Standard 6: Assessment intersects with posing, pausing, and probing only indirectly, but it does echo important purposes and characteristics of teacher formative assessment practice articulated in the higher levels of the posing, pausing, and probing construct maps. The INTASC "learning progressions for teachers" contend that teachers use "assessment" "flexibly" (to move/shift from level 1 to level 2) and that they "align assessment techniques to information needed...to improve instruction" (to move/shift from level 2 to level 3). This insight has echoes in the posing dimension of teacher formative assessment practice as expressed in level 4 of the posing construct map in particular. The posing construct map describes "flexible posing" at the "relational" level of teacher practice, in part: "Respondents...demonstrate flexibility in their questioning. They demonstrate an awareness of the variety of purposes of their questions and the need to match kinds of questions to specific purposes."

Lack of standards-based evidence concerning the construct(s) the "progressions" intend to measure. As previously mentioned, from a measurement/assessment

perspective, the INTASC "progressions," are a set of rubrics, not actual progressions. These rubrics—as is the case with all rubrics—are not construct maps. Rubrics should not be equated with the construct or constructs for which the rubrics are designed to play a role in assessing. Equating a rubric with a construct would be subscribing to a p-prim common to novice assessment designers (Duckor, Draney, Wilson, 2009).

This set of facts has implications when one compares the available evidence concerning the INTASC "learning progressions for teachers" to the 2014 testing standards and to the NRC's recommendations for educational assessments. Doing so highlights concerns about (a) lack of validation for the construct(s) the INTASC "progressions" purport to articulate and (b) the development process of the INTASC "learning progressions for teachers"—that it did not adhere to an essential recommendation regarding assessment design asserted by the NRC. This recommendation asserts that the cornerstone of the assessment design process should be a model of cognition concerning the construct under investigation (Pellegrino et al., 2001) and that this model should "always be based on empirical studies...and...ideally...provide a developmental perspective, showing typical ways in which learners progress toward competence" (Pellegrino et al., 2001, pp. 2-5).

Information about empirical studies upon which the INTASC "learning progressions for teachers" may or may not be based is not available. Literature about the "progressions" reports they are based on the INTASC Model Core Teaching Standards, not empirical studies of how teachers develop competence in constructs underlying the "progressions" (CCSSO, 2013).

When the *p-prim* "rubric equals construct" is effect, this has implications for the validity of the inferences, or score interpretations, drawn from the assessment process, even when the intended purpose of engaging in the assessment process is formative assessment. According to the 2014 *Standards* (AERA, APA, NCME), validation begins with an explicit statement of the proposed interpretation of test scores and rationale for the proposed use. This includes "specifying the construct the test [instrument or observation rubric] is intended to measure" (AERA, APA, NCME, 2014, p. 11). This specification describes in detail "the knowledge, skills, abilities, traits, interests, processes, competencies, or characteristics to be assessed" (p. 11). When a *p-prim* is in effect regarding specifying the construct, and construct maps are not used to articulate how learners—in this case teachers—progress in competency in the constructs under investigation; then attempts to provide a sound scientific basis for the inferences drawn from the assessment (the "proposed score interpretations" of the assessment), as efforts towards validation attempt to do, are compromised from the start.

Teaching for Robust Understanding (TRU) Framework. Schoenfeld's TRU Framework (2014), of which there is a mathematics-specific version and a general-application version, emerges from a sociocultural perspective on learning. TRU is the only well-known mathematics-specific framework to organize its rubrics by learning configuration in the classroom. TRU has different rubrics for individual work, small group work, student presentations, and one rubric for the whole class activities of launch, teacher exposition, and whole class discussion. The framework also has a "summary rubric."

Each of the five rubrics in the TRU Framework (one rubric for each configuration plus the summary rubric) describes three levels of teacher practice and classroom interaction. The levels are identified by the numerals *1*, *2*, and *3* in five "dimensions" for each rubric.

As was pointed out with uses of the term "dimension" regarding frameworks already reviewed in this chapter, use of the term "dimension" in this context is not synonymous with validated dimensionality of a construct. Again, this is a case of casual use of the term "dimension," not use of the term from a measurement or assessment perspective.

The five "dimensions" are (a) The Mathematics (or "The Content" in the general-application version of TRU), (b) Cognitive Demand, (c) Access to Mathematical Content (or "Equitable Access to Content" in the general-application version of TRU), (d) Agency, Ownership, and Identity, and (e) Uses of Assessment (or "Formative Assessment" in the general-application version of TRU).

Though the TRU Framework is relatively new (the rubrics debuted in 2014), the framework represents years of work. Schoenfeld details several iterations of the framework, including "disastrous" attempts (p. 610) in his insightful "Classroom Observations in Theory and Practice" published in *ZDM* in 2013.

The TRU math rubric was developed first as a research tool. Though the summary rubric seems straightforward, according to its creators, "actual use of the rubric [which involves using the subrubrics in the TRU framework] requires training" (Schoenfeld, 2015, p. 406). The "summary rubric" itself is not used for scoring (Schoenfeld, 2015). As previously mentioned, the TRU framework has subrubrics for whole-class instruction,

small-group work, student presentations, and individual student work.

Using TRU for scoring classroom observations "involves parsing classroom activities into a sequence of "episodes" of no more than 5 minutes each in duration, assigning scores to each episode using the relevant subrubric, and then computing a weighted average of scores" (Schoenfeld, 2015, p. 406). While not intended for scoring of episodes of classroom instruction, "the summary rubric," according to its creators and users, who have long and respected experience in research and professional development in the field, "does provide a clear sense of the kinds of classroom activities that will score high or low along each of the dimensions" (p. 406).

Relation of TRU with posing, pausing, and probing. Though the TRU framework does not explicitly reference questions, questioning, wait time, or probing, nor feature proxies for posing, pausing or probing, the framework speaks to issues closely related to some aspects of posing, pausing, and probing. It does this most strongly in the two of the five dimensions that are titled (a) "Agency, Authority, and Identity" (math-specific)/"Agency, Ownership, and Identity" (general) and (b) "Uses of Assessment" (math-specific)/ "Formative Assessment" (general). These two dimensions characterize, respectively, the extent to which classrooms provide students with "means for constructing positive disciplinary identities [students' identities] through presenting, discussing, and refining ideas" and demonstrate "responsiveness of the environment to student thinking" (Schoenfeld, 2016) [italics added].

Alignment with TRU "Agency, Ownership, and Identity" dimension. Though students' identities are not the anchoring concept of any of the construct maps in the P-P-

P Assessment, students' engagement and the extent to which they respond to poses, pauses, and probes are integral to the descriptions of teaching practice in all three construct maps. The probing construct map in particular underscores expectations that students talk and elaborate during lessons. This map articulates expectations of "extended episodes" of "on topic" student talk punctuated by probing (whether by the teacher or by other students) during whole class, small group, and one-on-one configurations. According to the probing construct map, instruction that demonstrates this is indicative of a level 4, "Contingent" or "relational" level probing item/task responses.

Alignment with TRU "Formative Assessment" dimension. Teachers' degree of responsiveness to student thinking is integrated into each of the three construct maps for the P-P-P Assessment. The more responsiveness to student thinking is demonstrated in a teacher's planning for, enacting, and reflecting on posing, pausing, or probing, the higher the respondent and responses to items are located on the P-P-P construct maps. In level 3 teaching practice on the TRU Summary Rubric in the "Uses of Assessment/Formative Assessment" dimension, "The teacher solicits student thinking and subsequent instruction responds to those ideas, by building on productive beginnings or addressing emerging misunderstandings." Tables 5, 6, and 7 show the content of the posing, pausing, and probing construct maps that parallels this notion. In each table, relevant content is organized by its connection to the planning, enactment, and reflection facets. Together, Tables 5, 6, and 7 underscore the significance of having descriptions of teacher proficiency in all three facets of practice.

Table 5: *Posing Construct Map Content Closely Related to TRU Framework's Level 3 on the Rubric for Dimension Five: "Uses of Assessment" (math-specific) or "Formative Assessment" (general)*

Facet	Posing
Planning	They plan questions that are contingent upon students' responses to these questions (e.g., they plan "hinge" questions and post-hinge question pathways for instruction). (Level 5, Extended Abstract) They plan a variety of questions designed to elicit a wide range of responses, including misconceptions and "unorthodox" responses. (Level 4, Relational)
Enactment	They tend to enact lessons that display several ways student responses can be used to further students' own and other students' learning regarding the lesson target. (Level 5, Extended Abstract) *Responses to items/tasks* indicate flexibility in posing to adjust to students' learning edges in real-time in relation to learning goals. (Level 5, Extended Abstract) Observation of teaching will likely show changing questioning strategies in response to student(s) response(s). (Level 4, relational)
Reflection	They are able to reflect on how questions posed functioned to elicit evidence of student understanding in relation to lesson objectives/target(s) of instruction. (Level 5, Extended Abstract) They are able to reflect on perceived effects of changing questions and/or questioning strategies. (Level 4, Relational)

Part of "solicit[ing] student thinking" and conducting "subsequent instruction" such that it "responds to those ideas"—*doing* formative assessment during lessons—in public school classrooms should mean striving to do so in an equitable manner. Pausing, and classroom routines to support pausing, can play a critical role in supporting the practice of more equitable in-class formative assessment. "Wait time"—pausing to provide students for time to process and reflect—has long been associated with positive effects on

students' and teachers' behaviors and attitudes in classrooms (Casteel & Stahl, 1973; Rowe, 1974; Stahl, 1990; Tobin, 1980, 1986, 1987). Among these positives, an increase in the number of students who volunteer responses has implications for equity, particularly equity of participation in the oral discourse of classrooms.

Table 6 presents the content of the pausing construct map most related to the TRU Framework's level 3 of the formative assessment/uses of assessment dimension. Note that integral to the content of the pausing construct presented in Table 6 is the notion of how pausing can support more equitable classroom participation and more systematic eliciting (sampling) of student thinking.

Table 6: *Pausing Construct Map Content Closely Related to TRU Framework's Level 3 on the Rubric for Dimension Five: "Uses of Assessment" (math-specific) or "Formative Assessment" (general)*

Facet	Pausing
Planning	They plan scaffolding for pausing that fosters student access to materials as needed to support thinking during pauses in relation to the learning goal. (Level 4, Relational)
Enactment	*Respondents who* adapt pausing procedures to a variety of cognitive and affective needs that are tied to demands of instruction. They can explain how their pausing moves benefit more systematic and equitable evidence gathering (e.g., pausing's role in increasing sample size), class/instructional participation, and decision making. (Level 5, Extended Abstract) *Responses to items/tasks* indicate pausing tailored to individual and group needs (e.g., ELs, students with 504 plans) and responsive to changing contexts. Pausing reflects purposeful attention in decision-making to student, context, and curriculum. (Level 5, Extended Abstract) Observation of teaching reveals contextualized use of "think time" based on explicit curricular challenges and/or student learning styles. (Level 5, Extended Abstract)
Reflection	They are able to reflect in detail with sophistication on the potential effects of alternative pausing moves for individual students, groups of students, and the whole class in relation to the learning target and in light of what they know about each individual and each group of learners. (Level 5, Extended Abstract)

The TRU Framework, without discussing pausing—or any other concrete/specific teacher practice—addresses supporting more equitable classroom participation and providing access to the lesson's content to a wide range of students from a different angle. It does so through its third and fourth "dimensions"—"Access to Mathematical Content/Access to Content" and "Agency, Authority, and Identity"/"Agency, Ownership, and Identity"— not through its fifth "dimension" "Uses of Assessment"/ "Formative Assessment."

Clearly, the FA Moves and TRU frameworks take differing approaches to describing and evaluating classroom interactions between teacher and students in terms of equity of participation, equity of access to lesson content, and equity of voice in classroom discourse. It remains to be seen what affordances and constraints might accompany these differing approaches, especially during attempts to improve teacher practice in these areas through intervention and targeted professional development, once teacher practice has been evaluated through assessment processes employing the frameworks. The content of the pausing construct map reveals that the conceptualization of teacher pausing it reflects is betting on improvements in teacher pausing to effect numerous positive changes related to equity. The pausing construct map reflects the hypothesis that greater sophistication in pausing will effect improvements in (a) equity of participation, (b) equity of access to lesson content, (c) equity of voice in classroom discourse, as well as (d) more equitable sampling of student thinking, and (e) improved teacher decision making.

Finally, Table 7 presents the content from the probing construct map that most closely

parallels the content of the TRU Framework's Level 3 on the rubric for dimension five, "Uses of Assessment"/ "Formative Assessment." Noteworthy in Table 7 is the content connected to the reflection facet of formative assessment practice. This portion of the probing construct map reveals an expectation of how teachers can learn to better orchestrate "subsequent instruction [that] responds to...student ideas, by building on productive beginnings or addressing emerging misunderstandings" (level 3 on the TRU Framework's rubric for dimension five, "Uses of Assessment"/ "Formative Assessment." Based on empirical evidence from my study, the probing construct map contends that before teachers can enact instruction that is productively responsive to student ideas, teachers need to be able to imagine scenarios where they do so and articulate ideas about who in their classes are likely to benefit most from such possible scenarios. This example highlights a critical way in which a construct map differs from a rubric. Much more so than a rubric does, a construct map suggests "how" a respondent can develop responses to items that suggest increasing proficiency in the latent variable of interest.

Table 7: *Probing Construct Map Content Closely Related to TRU Framework's Level 3 on the Rubric for Dimension Five: "Uses of Assessment" (math-specific) or "Formative Assessment" (general)*

Facet	Probing
Planning	They plan probes that should serve to elicit an intentional range of responses/performances in order to set up instructional decision making contingent upon what the probes elicit. They plan lessons that incorporate these probes and that can accommodate—or leverage—the implications of what the probes help make visible (e.g., modular, if-then, "flow chart"-like plans for lessons). (Level 4, Relational)
Enactment	They communicate concern for responding productively to student responses they cannot anticipate and can name strategies for doing so. (Level 5, Extended Abstract)
	Responses to item/tasks indicate pattern(s) to probing that include productive teacher responses to information newly elicited by probing and that is incorporated into further probing. (Level 5, Extended Abstract)
	Observation of teaching shows productive handling of surprise or "unorthodox" responses. Observation of teaching reveals that what probing elicits is progressively used to advance student responses toward the learning target. (Level 5, Extended Abstract)
	Respondents who adjust their probing methods or strategies according to incoming evidence ("evidence" that may or may not be gathered/processed systematically or strategically) and in light of the goal(s) regarding the learning target. (Level 4, Relational)
	They tend to enact probing that results in responses that get used by students and teacher. (Level 4, Relational)
Reflection	They are able to reflect in ways that articulate "next steps" that incorporate what probes were intended to or did reveal. That is, they are able to conjecture on possible, alternate post-probe pathways for instruction. They can speak to how improved probing might improve "options" for pathways for instruction and how this might benefit certain learners or groups of learners (e.g., "stuck" students, students holding certain misconceptions). (Level 3, Multistructural)

Validation of the TRU framework. Validity refers to the degree to which evidence and theory support proposed and actual uses of assessments (AERA, APA, NCME, 2014), including rubrics and frameworks intended for formative feedback tied to

classroom observations. How the two versions of the TRU framework, the mathematics specific version and the version for application to all content areas, are used in the field is presently still being determined. The TRU math rubric for classroom observations was first developed as a research tool and not intended for administrative use in evaluating teachers (Schoenfeld, 2014). Schoenfeld (2014) and the team responsible for the TRU framework "would much rather focus on working productively with teachers, as opposed to rating them" (p. 406).

In 2014, at the time Schoenfeld's commentary in *Educational Researcher*, "What Makes for Powerful Classrooms, and How Can We Support Teachers in Creating Them? A Story of Research and Practice, Productively Intertwined," was published, which included substantive information about the TRU math rubric, "validation of the rubric through research [was] in its very early stages" (p. 406). Presumably, since then validation work has continued. Information about the extent to which validation work has been carried out—and results of the work—however, have not yet been published. I expect that information will be published eventually, especially given the context of the framework's creation and development and its current use in multi-year projects with Delaware Mathematics Coalition and the "Math in Common Initiative" that involves 10 school districts in California, amongst other projects.

A goal of the Math in Common Initiative, a five-year project, is to shift mathematics instruction to be more aligned with the new CCSSM and to improve K-8 students' mathematics achievement. Classroom observations are being conducted in fall and spring with calibrated observers using the TRU framework. The data gathered will help

determine if teachers' instructional practices have shifted over the course of the five-year initiative.

As the TRU framework is used in the field, publication of validation work should follow. According to the 2014 *Standards for Educational and Psychological Testing* (AERA, APA, NCME):

> Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. (pp. 11-12)

The current purposes of using the rubrics for classroom observation in several school districts may be to serve in evaluating programs, not "rating teachers," but that does not suggest validation is not necessary. Nor, for that matter, is anyone on the TRU development team suggesting such is the case.

All the same, according to the *Standards*, validation is the joint responsibility of the test developer (or rubric for classroom observation developer) and the test user (or rubric user). It is also worth noting that according to the *Standards*, when a test (or classroom observation rubric) user proposes a use that differs from those supported by the test developer, the responsibility for providing validity evidence in support of that use is the responsibility of the user. Validation of the use of the TRU framework in the field is a recognized concern and likely will evolve related to the framework's use and its intersection with the educational environment. The educational environment, inevitably, will continue to respond to changing educational policies influenced by changing expectations for, and changing systems for, accountability in K-12 public education.

Construct Maps for the P-P-P Assessment

This final section of the chapter presents the development of the three construct maps critical to the design of the P-P-P Assessment. It discusses the timeline, rationale, and noteworthy specifics about three significant iterations of the construct maps in my study. First it explains the background and context for the hierarchical descriptions of "teacher knowledge, practice and reflection on formative assessment" that were the precursors of the first construct maps drafted for this study. Since these "pre-construct maps" were informed by Bloom's taxonomy and were not really construct maps, I refer to them as the "Bloom's Lite pre-construct maps." Next this chapter presents the first posing construct map, a version drafted in May 2016, as a representative from the set of three construct maps drafted before the study began. Finally, I describe some of the revisions that occurred between the first set of construct maps drafted in May 2016 and the close of the study and I present an exemplar from the resulting construct maps, the posing construct map.

Timeline of construct map development. Developing construct maps is an iterative process tied to the purposes of the assessment under design and development (Wilson, 2005). The posing, pausing, and probing construct maps underwent three significant revisions from their first nascent articulations in December 2013 through March, 2017. The study was proposed in September 2016, commenced in December 2016, and concluded in March 2017. The three iterations of the construct maps occurred

(1) before the study's proposal in September 2016,

(2) after receiving my committee's feedback during the proposal hearing; after

further, more targeted review of the literature; and before data collection commenced in December 2016; and

(3) after a significant amount of data had been collected and preliminary analysis had begun.

See Table 8 for a timeline of the three significant iterations of the construct maps employed in the design of the P-P-P Assessment.

Table 8: *Timeline of Iterations of the P-P-P Construct Maps*

Date	Revision	Reasons for Revisions
Dec. 2013	0	N/A (first drafts descriptions of "teacher knowledge, practice, and reflection on FA," aka as "Bloom's Lite pre-construct maps")
May 2016	1	● take from "Bloom's Lite" approach to SOLO taxonomy approach ● represent a "full" construct map (w/*respondents* side and *responses to items/tasks* side) ● focus solely on enactment (remove teacher planning and reflection)
Oct. 2016	2	● systematically add teacher planning and reflection ● incorporate Borko & Livingston's (1989) research on novice-expert teacher planning ● highlight (a) teacher anticipation, (b) leveraging what is elicited, (c) teacher decision making, and (d) equity
Feb. 2017	3	Based on empirical evidence: ● refine teacher reflection content ● foreground quality and range of student responses ● increase level of specificity concerning "variety" ● incorporate students' affective states ● add missed opportunities and alternative moves ● integrate components of lessons

Initial drafts of P-P-P "pre-construct" maps: the "Bloom's Lite" articulations.

Each draft of the P-P-P construct maps reflects a noteworthy stage in the honing of the purposes for their existence and use. The first articulations of teacher proficiency in the

posing, pausing, and probing dimensions of formative assessment hypothesized in this study—I refrain from calling these articulations construct maps, since they were not truly construct maps—began well before the beginnings of this dissertation study. Dr. Duckor and I sketched out hierarchical descriptions of teachers' knowledge, practice and reflection related to posing, pausing, and probing that we titled "Bloom's Lite: Categorizing Teachers' Knowledge, Practice, and Reflection of Formative Assessment."

We referred to these "pre-construct map" articulations as "Bloom's Lite" because the organizing principle we used to structure our descriptions of teachers' knowledge, practice, and reflection was informed by both the original and revised versions of the Taxonomy of Educational Objectives (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, & Wittrock, 2001), known popularly in the field as Bloom's Taxonomy and Bloom's Taxonomy Revised, or simply "Bloom's Revised."

Bloom's Taxonomy—original or revised—is a scheme for classifying educational goals, objectives, and standards that employs a cumulative hierarchical framework (Krathwohl, 2002). Achievement of the next, more complex skill or ability is expected to require achievement of the prior one on the taxonomy.

The structure of the original version of the taxonomy delineated six levels: knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al, 1956). Bloom's Revised articulates two "dimensions": a Knowledge dimension, of which there are four categories—factual, conceptual, procedural, and metacognitive—and a dimension encompassing Cognitive Processes. The six categories of the Cognitive

Process dimension in the Bloom's Revised are: remember, understand, apply, analyze, evaluate, and create (Anderson et al, 2001).

Figure 6 depicts a "Bloom's Lite" hierarchically-ordered description of teacher practice of formative assessment, a so-called *pre-construct map* for probing. This pre-construct map encompasses eight hypothesized levels. Note that while there are two columns associated with each level, the content in the columns does not make clear the purposes behind having two columns. A construct map, on the other hand, has a clear delineation between its characterizations of *respondents* and its characterizations of *responses to items/tasks* at each level. The pre-construct map in Figure 6 does not.

Creation/ Evaluation+	Can come up with likely, fitting and original FA solutions tailored to contextualized challenges, e.g. "What I tried wasn't working well enough; I've come up with possible solutions and why they are likely to work…"	Creates new practices, innovates
Evaluation	Has a way of figuring out whether it was effective; has a system for looking at its effectiveness I *see* what I'm doing, I know *why* I'm doing it, but it's *not working*: *how* can I…?	Strategically evaluates the effectiveness of the probing (by self and by students) in relationship to a goal (binning, deciding whether to offer an alternate explanation), e.g., "My probes would be more effective if I could word it such a way that…
Synthesis	Can provide explanation of why they're doing the application move; e.g. When asked why, they say, "I'm doing this... …to adjust pacing." … to check for misconceptions." …to find patterns of strengths and weaknesses…"	Uses probing in order to bin; Orchestrates probing (can scaffold students probing one another's thinking) in order to assist students' self-assessments

	…to build learning environment."	
Application+	Can exercise a variety of ways of doing; has a broader palette to choose from	Strives to make probes public and inviting; probes targeted at uncovering known common misconceptions; effective probes come from other students; norms/routines around probing (pair-share, writing, pausing; or ways to "soften" or normalize probing who feel it intrusive and uncomfortable)
Application	Can do; can ask probing questions	Asks probing questions to individual students (only), e.g: "What makes you say that?" "Because?" "How did come to that conclusion?" —metacognition, process, evidence, source, perspective, opinion/value judgment
Knowledge+	Can provide evidence that they know; e.g., can give criteria for probes and can give example probing questions	Recognizes/understands probing, purposes of probing are to elicit student thinking and make visible to teacher and learning community; probes tried may serve purpose "only" of seeing if students' conceptions conform to teacher's conceptions
Knowledge	Can give examples of probing	Knows follow up questions to student responses are valuable
Notions	Has fuzzy idea of probing	May be aware of opportunities to ask follow up questions; however, first responses "satisfy"/are sustained

Figure 6. Bloom's Lite pre-construct map for probing drafted before the study was proposed.

Inspection of Figure 6 reveals that the middle column is "teacher ability"-oriented and the right-hand column is "teacher action"-oriented, which is perhaps suggestive of an

attempt at articulating norm-referenced and criterion-referenced descriptions of practice. At the time of this pre-construct map's drafting, however, I was not aware of any real significance behind the two-column structure of this description of teachers' practices concerning probing.

Purposes of the "Bloom's Lite" pre-construct maps. As they were initially drafted, the "Bloom's Lite" articulations were not for the purpose of using them in the design of an assessment. I was a university supervisor of clinical practice for preservice teachers and I wanted to know if it were possible to coach my teacher candidates in ways that would "accelerate" their development. If "accelerating" their development were possible, what would help? My interest in fleshing out "Bloom's Lite" versions of posing, pausing, and probing, then, was to see how doing such an exercise could help me give better feedback to the preservice teachers I was supervising, feedback that I hoped would play a pivotal role in "accelerating" their development.

I was strongly oriented toward helping my teacher candidates meet their students where they were, become better listeners, gain practice in improvising during instruction, and maximize student engagement. I was also already experienced and skilled in facilitating evidence-based reflection with beginning and experienced teachers. Before becoming a university supervisor, I had worked closely with a wide range of developing teachers in a variety of contexts—from a half dozen student teachers in my own high school classroom, to preservice and inservice teachers attending Bay Area Writing Project workshops, to teachers seeking National Board Certification. I had taken a graduate-level course in the mentoring of preservice teachers, written a mentoring case

for that course, and participated in another professor's research project on mentoring new teachers of color.

I convey this background information to illustrate the position from which I began my work on the P-P-P construct maps. If construct maps, as Lehrer and Schauble (2009) contend—in agreement with Wilson (2009) — "distill...proposals...about what is worth knowing (as well as what may be considered less necessary), commitments about the forms of evidence that are most relevant to learning, and informed judgments about trajectories of prospective development" (p. 734); then, even if I did not realize it at the time, I was in a strong position to begin the work of drafting construct maps, ones that I hope will serve the larger purpose of informing a developing teacher learning progression in formative assessment.

First versions of construct maps: from Bloom's to Biggs and Collis. In May 2016 I took some of the content in the "Bloom's Lite" articulations of teachers' knowledge, practice, and reflection and drafted my first construct maps. May 2016 marked the first time I described the variation I expected to see in teacher enactment of posing, pausing, and probing with respect to persons and items.

I did so for five levels of proficiency since I was employing the Structure of Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982) to the project of conceptualizing the structure of teacher learning of formative assessment. As a general theoretical framework that has been used in many assessment contexts as a way to get started in constructing *outcome spaces* for tasks related to cognition, the SOLO taxonomy is also a sound choice to use as a basis for making explicit hypotheses about how

proficiency in formative assessment progresses in teachers, i.e., drafting a construct map (Wilson, 2005).

Figure 7 displays the May 2016 draft of the posing construct map, my first attempt at employing the SOLO taxonomy in drafting a construct map. I drafted the pausing and probing construct maps similarly, employing the SOLO taxonomy. Following Figure 7, I describe salient characteristics of that set of construct maps. These characteristics largely resulted from my focus at the time I was creating them. While drafting them I was focused on lesson enactment to the exclusion of attention on planning and reflection.

High	*Respondents who* integrate relevant features of the context for learning with multiple important purposes for questions (e.g., promoting making thinking visible or meta-cognition). They pose questions that size up the context for learning in ways that reflect knowledge of students' development, interests, needs for learning target(s) and present understandings. They pose questions that relate to the lesson and the unit plan and larger essential questions/big ideas of the discipline.	Integrative posing 5	*Responses to items* indicate flexibility in posing to adjust to students' learning edges in real-time in relation to learning goals. Questions posed leverage a range of student responses in ways that elicit evidence of having furthered students' present understandings in relation to the lesson target and/or essential question/big idea of the discipline. Show that respondent has anticipated student pit stops and bottlenecks typical of learning progression of concept/skill/understanding.
	Respondents who demonstrate flexibility in their questioning. They demonstrate an awareness of the variety of purposes of their questions and the need to match kinds of questions to specific purposes.	Flexible posing 4	*Responses to items* indicate posing of *how* and *why* questions and questions from a mix of DOK levels. Teacher may change questioning strategy in response to student(s) response(s).
	*Respondents who*se purposes for questioning seem to be to get students to say what they are thinking (rather than eliciting from students a range of responses, including unknown responses, responses surprising to the teacher).	Constrained posing 3	*Responses to items* indicate posing a high percentage of, or posing only, *what/when/where*, fact recall, and lower-level questions (on Webb's DOK, Bloom's taxonomies etc.). Questions posed connect to learning target. Questions posed do not elicit a wide range of responses.
	Respondents who demonstrate through their questioning a primary focus on orchestrating student behavior, not necessarily learning. They may not be able to make student thinking visible through questions they pose.	Posing to manage 2	*Responses to items* indicate posing to manage/control students, e.g., "Do you have a pencil? Are your books open to page 39?"
Low	*Respondents who* give directions to students and whose actions can be interpreted as attempting to pour content into students' minds without eliciting from the students where their current understandings are.	Pre-posing 1	*Responses to items* indicate no questions are posed by the teacher.

Figure 7. May 2016 draft of posing construct map.

A focus only on enactment. In Spring 2016, my understanding of the purposes of the assessment I was embarking on designing caused me to omit from my May 2016

construct maps references to planning and reflection. I attempted to focus the May 2016 construct maps on what was directly observable during lessons. I did not have a clear sense of how I would collect evidence of teacher reflection on practice.

Limited hypotheses about purposes. This resulted in few statements amongst the May 2016 construct maps that addressed teachers' purposes behind actions representing instantiations of posing, pausing, and probing moves. This was regardless of whether or not such purposes were intended by teachers ahead of enactment (planned purposes) or not. Later versions of the construct maps distinguished teachers' intended purposes—as stated by P-P-P Assessment respondents, for example, in lesson planning documents collected during the study, or as reported during reflection catalyzed by the video-stimulated recall (VSR) protocol during the "reflection session" interview—from purposes teachers' moves implied during enactment. Later versions of the construct maps made these distinctions because it became apparent that paying attention to teachers' stated intended purposes as compared to paying attention only to their "demonstrated" (or enacted) purposes was helpful to locating respondents and responses to items on the continua of practice hypothesized in the study.

Second drafts of P-P-P construct maps motivated by study proposal hearing feedback and further literature review. The purposes behind the next set of revisions to the P-P-P construct maps were two-fold: to follow through on feedback about the construct maps received during the study proposal hearing and to incorporate further—targeted—literature review into the maps. Both were related to systematically adding hypotheses about teacher planning and reflection into the construct maps. During my process of composing my study proposal I came to a new understanding of what the cycle

of planning a lesson, enacting the lesson, and reflecting on the enacted lesson would mean for my study and for the design of my assessment. There was no way I could move forward without explicitly incorporating planning and reflection into my heretofore enactment-focused-only construct maps.

Figure 8 presents two versions of the multistructural (3) level of the posing construct map: versions of it before and after the revisions motivated by feedback I received during the hearing of my study proposal. These revisions are representative of the process I carried out for every level of all three construct maps. The revisions, which are mostly additions of content, are highlighted in Figure 8.

May 2016		
Respondents whose purposes for questioning seem to be to get students to say what they are thinking (rather than to elicit from students a range of responses, including unknown responses, responses surprising to the teacher).	Constrained Posing (Multistructural) 3	Responses to items indicate posing a high percentage of, or posing only, *what/when/where* fact recall, and lower-level questions (on Webb's DOK, Bloom's taxonomies, etc.). Questions posed connect to learning target. Questions posed do not elicit a wide range of responses.

October 2016		
*Respondents who*se purposes for questioning seem to be to get students to say what the respondent-as-teacher is thinking (rather than eliciting from students a range of responses, including unknown responses, responses surprising to the teacher). They plan questions they consider checks for understanding of the lesson's objective. Then tend to enact lessons with high percentages of close-ended questions. They tend to enact lessons in which scenarios arise where students are expected to guess what the teacher is thinking, even when doing so appears more a hindrance to than a help regarding students' advancement toward the learning target. They are able to reflect on several aims of improving posing. Their reflection includes specific suggestions for alternate poses to try.	Constrained Posing (Multistructural) 3	*Responses to item/tasks* indicate posing a high percentage of, or posing only, *what/when/where*, fact recall, and lower-level questions (on Webb's DOK, Bloom's taxonomies etc.). Questions planned connect to learning target. Observation of teaching shows questions posed as checks for understanding procedures and concepts tied to the learning target. Observation of teaching shows questions that seek to elicit students' prior knowledge. Observation of teaching reveals questions posed seldom elicit a wide range of responses.

Figure 8. Revisions (highlighted) to the multistructural level of the posing construct map that are representative of revisions carried out with all the levels of all three construct maps in the study.

Figure 8 presents the revisions for multistructural (3) level. The levels between the

extremes on a construct map—in my case, levels 2, 3, 4—are of particular significance. During efforts to map a variable of interest—and during tests of the functioning of a construct map within a particular assessment context—expert assessment designers employing Wilson's "building blocks" approach to design and develop an assessment are "intensely interested in the levels in between...'experts' and 'novices' (Duckor et al, 2009, p. 302). They want their construct maps to identify and communicate indicators of progress that can be used to reliably, and meaningfully, locate respondents and responses to items/tasks amongst these middle levels. This is in line with an assessment that is being designed for the purposes of evaluating teachers' instructional practice, an assessment whose aim is also to be able to generate "next steps" feedback helpful to teachers in their efforts to further develop their proficiency in formative assessment.

A construct map's capacity to reliably and meaningfully indicate different locations for respondents and responses to items/tasks influences its potential to function well in helping stakeholders use the score interpretations derived from the assessment based upon that construct map. Instructional coaches, professional developers, administrators, teachers themselves, and other educational stakeholders interested in helping develop teachers' proficiency in formative assessment, therefore, value what accurate descriptions of the "messy middle" as Gotwals and Songer have called it (2010, p. 277), can serve.

Incorporating research on novice-expert lesson planning. Borko and Livingston's (1989) research on the differences between novice and expert teacher lesson planning especially informed this second significant revision of the P-P-P construct maps. Superfine's (2008) work on experienced teachers' planning processes in the context of a

reform curriculum did too. Specifically, additions to the construct maps that referenced the notion of teachers planning questions and probes that reflect attempts to balance content-centered instruction with student-centered instruction were inspired by this further literature review of mathematics teachers' lesson planning. Additions to the construct maps that referenced challenges in prioritizing the purposes of questions teacher planned came from this more targeted review of the literature too. The research clearly communicated that novice teachers experienced difficulty deciding what mathematics content was most important to question and probe.

Other revisions. Further revisions to the construct maps fell under four themes: (a) teacher anticipation, (b) leveraging student responses, (c) teacher decision making, and (d) equity. Adding content to the P-P-P construct maps that can be categorized by these themes served to expand the level of detail concerning what the maps together were hypothesizing about formative assessment more broadly. In making this set of revisions, I used what I gleaned from the research literature that was relevant to these themes. The work of Jacobs, Lamb, and Philipp (2010) and Borko, Roberts, and Shavelson (2008) especially influenced these revisions.

Revisions to P-P-P construct maps motivated by study data. The P-P-P Assessment's construct maps strive to treat teacher planning for, enactment of, and reflection on posing, pausing, and probing systematically and meaningfully. Teacher responses to the "FA Moves Lesson Planning Tool," a lesson planning template, played an important role in eliciting evidence of teacher planning. The template was presented to

subjects during the first interview session. A response to the lesson planning template is presented in Figure 10 on pages 124-127 in chapter 3.

Lesson planning template responses-inspired revisions. Responses to the common lesson planning template presented to subjects, the "FA Moves Lesson Planning Tool," motivated a specific revision across all three construct maps. After considering the empirical evidence, I integrated explicit and implicit references to "lesson components" into the P-P-P Assessment's construct maps. For example, I added explicit references to components of a lesson to levels 4 and 2 of the probing construct map. For respondents I was locating at the relational level (4), I added, "They plan for student-to-student probing to occur in some components of their lessons." For level 2, or Task-focused, unistructural-level probing, I added "[Respondents] tend to enact most of their probing in one component/portion of the lesson."

An "implicit reference" to a lesson component in the P-P-P construct maps is content that does use word "component" but nonetheless speaks to the notion of a lesson being able to be conceptually partitioned into smaller parts. By this definition, I added an implicit reference to the components of lesson—and the use of pausing moves throughout the entirety, or near entirety of a lesson—when I added to the respondents side of the pausing construct map for level 4, relational pausing, "[Respondents] tend to enact a mix of "quiet", "noisy", "active", "still", "individual" and "group" pausing moves and routines *throughout a lesson* to fit lesson goals, pacing, and learners' needs [italics added]. I added an implicit reference to the components of a lesson in relation to probing by adding to the description of level 3, multistructural, targeted probing, "[Respondents

plan specific probes for different point in the lesson and the probes reveal the teacher's expectations of progression."

Another such implicit reference I added to the posing construct map at this time, motivated by my initial consideration of responses to the "FA Moves Lesson Planning Tool," was "[Respondents at this level] plan carefully sequenced repetition of key questions. I asserted that evidence of this kind of planning would suggest level 4, relational posing. While this revision does not use the word "component," it references the concept indirectly. Since Bruner's introduction of "spiral curriculum" (1960), instruction that features intentional, skillful, and recursive posing of the same, and worthy, question over time—even within one lesson—has been recognized as an effective teaching practice to promote student learning. Teachers in the study who wrote out poses and probes word for word in response to the "FA Moves Lesson Planning Tool," and did so in more than one lesson component, motivated this addition.

Moreover, all these revisions were inspired by empirical evidence demonstrating a pattern: the more lesson components in which any FA move was referenced in a response to the "FA Moves Lesson Planning Tool," the stronger the likelihood of the respondent demonstrating a higher enactment proficiency than respondents whose lesson components lacked references to FA moves. This held for whatever lesson components teachers identified, since the "FA Moves Lesson Planning Tool" left it open for teachers to decide what was meant by "lesson component."

Refinement of teacher reflection content. The study data induced me to refine the content in the construct maps that referenced teacher reflection. Preliminary analysis of

the reflection evidence of five teachers inspired me to add the concepts of "missed opportunities" and "alternative moves" in the construct maps. This reflected a new-found understanding on my part, as an assessment designer, that I was valuing "multiplicity and specificity" in teacher reflection on practice. That is to say, a teacher's ability to proffer multiple explanations for student behavior, take multiple points of view regarding practice/enactment, and express multi-faceted purposes behind actions during reflection needed to be recognized across the construct maps and sensibly incorporated.

During reflection, half of the teachers in the study spoke of using, or wishing to use, various schemes for improving their questioning. One teacher reported a next step to try to "get better" would be to "use a questioning bank." Another reported having relied on Bloom's Taxonomy while she was interning in another school district, but who had not been using that "technique" since teaching with a new mathematics curriculum. Another teacher reflected on the strengths of her habit and system of posting—and keeping available throughout the week—the "essential question for the lesson." My revisions to the construct maps reflect this evidence.

Version used to locate respondent proficiencies. Figure 9 presents an exemplar construct map, the map for posing, from the three construct maps used to locate respondents' proficiencies in posing, pausing, and probing based on evidence generated from their engagement with the P-P-P Assessment. The three construct maps (see appendix A), and respondents' performances in comparison to them, were also used to create the profiles of practice featured in chapter 5.

High	*Respondents who* integrate relevant features of the context for learning with multiple important purposes for questions (e.g., promoting meta-cognition). They pose questions that size up the context for learning in ways that reflect knowledge of students' development, interests, needs re: learning target(s), and present understandings. They pose questions that relate to the lesson and the unit plan and larger essential questions/big ideas of the discipline. They plan questions that reveal explicit anticipation of where students may/are likely to get stuck or have misconceptions. They plan questions that serve to provide evidence in helping teachers decide which of a few to several specific (and expressed) decisions they might make that are contingent upon students' responses to these questions (e.g., they plan "hinge" questions and post-hinge question pathways for instruction). They plan in ways that encourage student questions to be springboards for discussion. They plan questions that reflect a balance between content-centered instruction with student-centered instruction. They tend to enact lessons that display several ways student responses can be used to further students' own and other students' learning regarding the lesson target. They tend to enact lessons that feature questions that reflect a sensible balance in addressing a variety of learners' needs. They are able to reflect on how questions posed functioned to elicit evidence of student understanding in relation to lesson objectives/target(s) of instruction.	Integrative posing (Extended Abstract) 5	*Responses to items/tasks* indicate flexibility in posing to adjust to students' learning edges in real-time in relation to learning goals. Questions posed leverage a range of student responses (including student questions) in ways that elicit evidence of having furthered students' present understandings in relation to the lesson target and/or essential question/big idea of the discipline. Responses to items/tasks show that respondent has anticipated student pit stops and bottlenecks typical of learning progression of concept/skill/understanding. Observation of teaching shows student responses being used in a variety of ways, including changing the direction of the lesson and/or pausing an activity.
	Respondents who demonstrate flexibility in their questioning. They demonstrate an awareness of the variety of purposes of their questions and the need to match kinds of	Flexible posing (Relational) 4	*Responses to items/tasks* indicate posing of *how* and *why* questions and questions from a mix of Webb's DOK or other taxonomic levels (e.g. Bloom's, Costa's).

	questions to specific purposes. They plan a variety of questions designed to elicit a wide range of responses, including misconceptions and "unorthodox" responses. They plan carefully sequenced repetition of key questions. They plan supports/scaffolds for questions. Then tend to enact lessons in which activities and pacing clearly reflect teacher decisions that are contingent upon student responses to questions posed about the learning target. They are able to reflect on perceived effects of changing questions and/or questioning strategies. They are able to suggest several "next steps" likely to support improved posing. They do so from many perspectives and with specificity.		Observation of teaching will likely show changing questioning strategies in response to student(s) response(s). Observation of teaching may show students playing significant roles in posing questions. Observation of teaching will show many questions that serve to highlight connecting students' prior knowledge and experiences with present efforts to engage with and "reach" the learning target.
	*Respondents who*se purposes for questioning seem to be to get students to say what the respondent-as-teacher is thinking (rather than eliciting from students a range of responses, including unknown responses, responses surprising to the teacher). They plan questions they consider checks for understanding of the lesson's objective. Then tend to enact lessons with high percentages of close-ended questions. They tend to enact lessons in which scenarios arise where students are expected to guess what the teacher is thinking, even when doing so appears more a hindrance to than a help regarding students' advancement toward the learning target. They are able to reflect on several aims of improving posing. Their reflection includes specific suggestions for alternate poses to try.	Constrained posing (Multistructural) 3	*Responses to item/tasks* indicate posing a high percentage of, or posing only, *what/when/where*, fact recall, and lower-level questions (on Webb's DOK, Bloom's taxonomies etc.). Questions planned connect to learning target. Observation of teaching shows questions posed as checks for understanding procedures and concepts tied to the learning target. Observation of teaching shows questions that seek to elicit students' prior knowledge. Observation of teaching reveals questions posed seldom elicit a wide range of responses.
	Respondents who demonstrate through their questioning a primary focus on orchestrating student	Posing to manage (Unistructural)	*Responses to items/tasks* indicate posing to manage/control students, e.g., "Do you have a pencil? Are your

Low	behavior, not necessarily learning (activity-based posing). They may not be able to make student thinking visible through questions they pose. They plan questions that do not reveal clear priorities in the purposes of posing questions. As they plan, they experience challenges in deciding what content is most important to ask about and when. They tend to enact teacher-centered lessons that do not reflect an underlying pedagogical structure dependent upon student responses to curricular content. They are able to reflect on benefits that might accrue from using a questioning scheme.	2	books open to page 39?" Planned questions do not express recognizable coherence or organizing principle. Observation of teaching and questions reveal imbalance of focus between activity/behavior and learning target.
	Respondents who give directions to students and whose actions can be interpreted as attempting to pour content into students' minds without eliciting from the students where their current understandings are. They may plan questions not well-crafted to elicit evidence of student understanding in relation to instructional goal(s). They tend to enact lessons that do not invite or incorporate students' prior knowledge. They are able to "reflect" through descriptions of their instruction that do not push to analysis.	Pre-posing (Prestructural) 1	*Responses to items/tasks* indicate no questions aligned with lesson target are posed by the teacher. Planned questions may or may not align with lesson target(s). Observation of teaching may show random or arbitrary questions.

Figure 9. The posing construct map for the P-P-P Assessment.

Requirements for sound assessment design. It is critical for an assessment designer to move beyond the p-prim that "the latent construct is the items." It is also critical for an assessment designer to move beyond simply—and incorrectly—equating the latent construct with the responses to items, as some novice assessment designers will do

(Duckor, 2005; Duckor, Draney, & Wilson, 2009; Wilson, 2005). Being able to come up with an items design for an assessment, an items design that will function well to help determine respondents' degrees of possession of the latent construct, depends on moving beyond these p-prims. In the next chapter, chapter 3, Items Design, we see shall see why and how. The inferential nature of assessment, the fallibility of observations, and the need for bodies of evidence from several different sources to support arguments for validity concerning an assessment are all intricately linked to the hypotheses of learning progression embodied in a construct map. Wilson's Constructing Measures framework has been explicitly designed to help assessment designers meaningfully integrate the complexity of these notions, concepts, and truths as they use the building blocks—starting with the construct maps building block—to design and develop an assessment.

"Building blocks" mirror principles of assessment advocated by NRC. The National Research Council committee (2001), whose expertise resulted in the seminal publication Knowing What Students Know: The Science and Design of Educational Assessment and the assessment triangle introduced in chapter 1, argues that a model of cognition and learning should serve as the cornerstone of the assessment design process" (p. 2). Furthermore, this model of cognition and learning should always be based on empirical studies of learners in a domain" and "Ideally, the model will also...show typical ways in which learners progress toward competence" (pp. 2-5). By employing Wilson's Constructing Measures approach and by creating the posing, pausing, and probing construct maps based on empirical studies of teachers (e.g., Borko, Superfine, etc.), in the design of the P-P-P Assessment I am following the recommendations of experts in

educational assessment and the principles and logic of assessment represented in the NRC's assessment triangle. This in turn increases the likelihood of the P-P-P Assessment, when used as intended, being well positioned to be able to meet the 2014 testing standards as needed.

Chapter 3: Items Design

This chapter defines the concept, *items design*, outlines characteristics of good performance tasks, addresses why they are used in evaluating the practices of teachers, and presents the items design of the P-P-P Assessment. The chapter begins by defining the concept, *items design,* and introducing the *performance task* as a valued item type within the population of items—or *item pool*—of an assessment.

Since the P-P-P Assessment relies on three performance tasks within its items design to generate construct-relevant and construct-representative responses, the chapter's second section describes the characteristics of performance tasks as an item type. It presents characteristics that contribute to effective functioning of performance tasks within a performance assessment. It also outlines features of performance assessments that have been widely used to evaluate both beginning and experienced teachers' instructional practices. A body of research has shown that the use of these performance assessments benefits the learning and practices of the teachers who participate in these kinds of activities. Overall the section implicitly explains why performance tasks are used when an assessment's aim is evaluating teacher practice.

Finally, the chapter's third section presents the items design of the P-P-P Assessment and accompanying rationale for its design. It features a timeline of the items of the P-P-P Assessment in the sequence in which they were presented to subjects. This is followed by the results of the task analyses of two key items: a planning-focused item and a reflecting-focused item. These are the lesson planning template known as the "FA Moves Lesson Planning Tool" and the protocol used to stimulate teachers' reflection on video

evidence of lesson enactment. The goal of this latter item, the video-stimulated recall (VSR) protocol, was to generate responses relevant to and representative of the reflecting facet of the posing, pausing, and probing dimensions of FA practice. The chapter closes with the table of specifications for the P-P-P Assessment.

Definition

An *items design* can be understood as a description of all the items—the prompts, tasks, questions, and follow-up questions, for example—intended for use in an assessment (Wilson, 2005, p. 44). The purpose of an items design is to stimulate and structure opportunities for observing respondents' knowledge, cognitive processes, and performance during the assessment in order to make inferences about respondents' "possession" of the targeted construct—or constructs—of interest. Evidence of a theoretical construct cannot be directly observed (Wilson, 2005).

The only way "possession" of "amounts" (e.g., "less" or "more") of a construct can be inferred is through interpretation of observed responses. Therefore, assessment designers must think of ways by which the theoretical construct can be manifested in real-world situations. Assessment designers carefully construct an items design, aiming for the items design to function well in strategically and systematically eliciting construct-representative and construct-relevant responses from assessment takers.

They do so in a manner such that the responses can be analyzed and verified with reference to other evidence, such as, for example, evidence of assessment takers' *response processes*[1]. The aim in working this way and gathering such other, additional

[1]Evidence based on *response processes* is one of five sources of validity evidence recognized in the 2014

evidence beyond the responses to the items themselves is to support arguments for validity. Items designs should serve to establish sufficient levels of validity and reliability for the instrument [or assessment] (Wilson, 2005, p. 44).

Decisions. An items design results from many decisions the assessment designer makes about how to represent the construct. Inherent to representing the construct through items chosen and/or created for an assessment's items design is deciding "how to stratify the 'space' of items...and then sample from those strata" (Wilson, 2005, p.45). An assessment designer's choices should be explicitly and soundly related to the other building blocks in the Constructing Measures (CM) framework: the construct maps, outcome space, and measurement model building blocks (Duckor et al., 2009; Wilson, 2005).

On the whole, choices made about an items design should function to strengthen inferential links between specific aspects of the measurement framework (Duckor et al., 2009). When the choices made do strengthen inferential links, the design, development, and functioning of the assessment are better-positioned to align with recommendations for educational assessments made by the National Research Council (2001) (e.g., that the model of *cognition*, modes of *observation*, and methods of *interpretation*—the three vertices of the assessment triangle—work in synchrony to support the meaningfulness of inferences drawn from the assessment) than when the choices do not. When these conditions are met, it follows that, when used as intended, the assessment will also be

---

Standards for Educational and Psychological Testing. It involves investigations into the fit between the construct and the nature of the performance or response actually engaged in by test takers (AERA, APA, NCME, p. 15). Generally this comes from analyses of individual responses. Techniques include questioning test takers from various groups making up the intended test-taking population about their performance strategies or responses to particular items.

better positioned to meet the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

As with other decisions related to designing an assessment, choices involving trade-offs are unavoidable. It is critical that an assessment designer be aware of the implications of decisions made regarding an items design, both potential implications and implications discovered and confirmed by analyses of item responses and analyses of additional empirical evidence gathered from assessment takers (e.g., response processes validity evidence). Experts in the domain of items design demonstrate such awareness by being able to indicate where and when a particular items design is likely to strengthen/weaken inferential links between specific aspects of the (CM) measurement framework (Duckor et al., 2009).

Such experts may use analyses of data to inform revisions to construct maps, the items design, the outcome space, and even the measurement model (Duckor et al., 2009) since designing and developing an assessment is an iterative process. As the NRC recommends, assessment developers will look for "mismatches" and "refine the elements...to achieve consistency" (Pellegrino et al., 2001, p. 51) amongst all the elements of the assessment triangle. As previously mentioned, the building blocks of the CM framework are proxies for the vertices of the NRC assessment triangle.

Challenges. The challenges of constructing an items design are many. In conditions where assessment designers wish to represent a wide range of contexts for the assessment (which is the case with the larger, longer-range aims and goals of the P-P-P Assessment), assessment designers recognize that sampling more of the content of a construct is better.

This "more sampling" is generally accomplished by having more, rather than fewer, items in an items design. When an items design works to generate more "bits" of information about how a respondent stands with respect to the construct, this gives greater accuracy to an assessment (Wilson, 2005, p. 44).

At the same time, the item *formats* of the items in the "item pool" of an assessment—the item "pool" is the population of items in an assessment—need to be "sufficiently complex to prompt responses that are rich enough to stand the sorts of interpretations that the [assessment designer] wishes to make with the [assessment]" (Wilson, 2005, p. 44). Too many simplistic item formats (e.g., Likert scales or fixed-choice/multiple choice items) that generate responses that yield little insight into the respondents' possession of the construct within an items design will weaken its functioning. Taking the assessment will result in insufficient evidence to support inferences about complex, multidimensional phenomena.

On the other hand, no single "true task," no matter how "authentic" an item it may be, "will supply the mother lode of evidence about the construct" (Wilson, 2005, p. 44). The tendency to search for—or to expend effort to create—"the *one best* task" is common among novice assessment designers (Braun & Mislevy, 2005; Wilson, 2005) and plays into not being able to establish sufficient evidence supporting validity and reliability for an assessment. A mix of item formats, tasks, and observations within an items design can yield better results (Duckor, Draney, & Wilson, 2009). For example, an assessment designer might include items that spur self-report along with items, such as performance tasks, that spur respondent actions that can be video recorded. Making such items design

choices should take into account the purposes of the assessment and knowledge of the theorized construct. These are not decisions neutral to the idea of the construct (Wilson, 2005), nor should they be.

Amount of pre-specification of item formats. Some of the challenges inherent in the decision-making processes of assessment design and development related to items design pertain to the concept of pre-specification of item formats. An important way to characterize, reflect on, and analyze the items and item formats populating an items design, especially during the development of items, is to consider the differing amounts of pre-specification of each item format within an assessment's items design. Pre-specification is the degree to which the results from the use of the instrument are developed before the instrument [assessment] is administered to a respondent (Wilson, 2005). The more that is pre-specified, the less that has to be done after the response has been made.

As a characteristic of item format, pre-specification is important more than just as a way to differentiate amongst item formats in the "pool of items" for an assessment. For the assessment designer, knowledge of the amount of pre-specification of item formats can help the designer make strategic choices concerning the amounts of pre-specification of item formats utilized in an items design. The goal is to achieve the optimum amount of pre-specification *during the development* of items for an assessment. Generally, designers start the development of items with low amounts of pre-specification and proceed to greater amounts of pre-specification until the optimum amount is reached (Wilson, 2005). Use of knowledge of pre-specification can help an assessment designer strategically meet

the challenges inherent to designing and developing an assessment to meet its intended purposes.

Imposed limitations. Some assessment design challenges pertain to imposed limitations. An ever-present tension in the decision-making process of creating an items design for an assessment is meeting the aims of the items design—and the purposes of the assessment—within the imposed time and cost limitations of the assessment context. In the case of the P-P-P Assessment, the imposed time limit was due to agreements reached with supervising administrators at the study sites and with study participants.

Observation time for the P-P-P Assessment was limited to what could be stimulated and gathered from respondents during two online surveys and three in-person sessions. Each in-person session was 90 minutes long. The three in-person sessions were (a) an intake/planning interview, (b) the lesson enactment (which was video recorded), and (c) a lesson reflection interview. The two online surveys were intended to take respondents approximately 20-25 and 5-10 minutes respectively to complete.

Qualitative levels targeted by items. Another important decision an assessment designer makes when drafting and honing an items design for an assessment is deciding which qualitative levels will be targeted by the item pool—and by specific items within that pool. The designer seeks coverage of the construct. Since the purpose of an items design is to elicit responses from assessment takers that represent the construct, and since the items design is linked to the outcome space, an items design that does not function to elicit important evidence about the qualitative levels of performance in the responses of assessment takers is not doing its job well.

A goal is to be able to locate each respondent and each respondent's responses on the relevant construct map or maps. An items design reflects the fact that an individual item should always target at least two qualitative levels; otherwise the item would not be useful in its goals of serving (a) to learn about the construct and (b) to locate each respondent and each respondent's response—by item—on the construct map (Wilson, 2005, p.45). Critical to supporting the validity argument for an assessment is that these qualitative levels—which are at first hypotheses, even if they are research-supported hypotheses (as they were in the design of the P-P-P Assessment)—are adjusted, or not adjusted, according to empirical evidence gathered during the assessment development process.

An individual item can be designed to generate responses that span only two qualitative levels. If so, deciding upon which two levels is crucial for achieving "balance" —or, perhaps, strategic and intentional "imbalance"—in the (overall) items design of an assessment. An individual item can also be designed to generate responses that span more qualitative levels than that, up to the maximum number of levels in the construct (Wilson, 2005).

Performance tasks generate responses spanning several qualitative levels. As a type of item, a performance task is often valued by assessment designers for its characteristic ability to generate responses spanning several qualitative levels. Assessment designers do not typically include a performance task or tasks in an items design in order to generate responses spanning only two qualitative levels. A performance task is also valued by assessment designers as an open-ended item format, in

contrast to fixed-response format items. Open-ended format items, which include interviews, essays, proofs, and multi-part products, do not expect a single, correct answer in response. Open-ended item formats expect respondents to use more than one approach while responding to the item. The next section addresses several characteristics of performance tasks. This is because as an open-ended item format, the performance task served as a central feature of the P-P-P Assessment's items design. Specifically, the P-P-P Assessment relied on three performance tasks in its items design: one aligned with planning, another focused on enactment, and third that targeted reflection.

Performance Tasks as Items

Performance assessments are "assessments for which the test taker actually demonstrates the skills the test is intended to measure by doing *tasks* that require those skills" AERA, APA, NCME, 2014, p. 221) [emphasis added]. A *performance task* is an assessment activity that requires those being assessed to demonstrate their achievement by producing extended spoken or written responses, engaging in group or individual activities, or creating a specific product (Brookhart & Nitko, 2015).

In the context of assessment of K-12 students, performance tasks typically require students to directly demonstrate their achievement of learning objectives when the learning objectives are beyond simple recall of facts or comprehension of content. As an item type, performance tasks are best matched to situations requiring students' application of knowledge and skills in context. Assessing complex student learning objectives calls for complex assessment (Arter, 1998).

Assessment designers strive to create performance tasks that will be effective in

eliciting sufficient evidence about complex, multidimensional phenomena. While recognizing that no *single* task, no matter how "authentic" "alternative" or "rich," can provide all the evidence necessary for providing sufficient levels of reliability and validity for the assessment (Wilson, 2005), designers strive to include effective performance tasks in their items designs. A performance task should function to elicit evidence of an assessment taker's "possession" of levels of the construct or constructs under investigation.

Characteristics of good and effective performance tasks. Performance assessment in general encompasses several types of performance assessment techniques. Techniques include (a) structured, on-demand tasks for individuals, groups or both (such as demonstrations), (b) naturally occurring performance tasks (such as when assessors observe those being assessed in natural settings in class, on the playground, or at home), (c) longer-term projects for individual students, groups, or both (such as oral presentations), and (d) simulations (Brookhart & Nitko, 2015). Affordances and constraints inherently accompany each type. An affordance of simulations, for example, is that the conditions under which they occur are—typically—tightly controlled. They offer more standardized conditions than conditions present when assessment takers engage other types of performance tasks, especially compared to performance tasks where assessment takers can choose the setting in which they perform the task.

While performance tasks of many different types exist, not all performance tasks are good or effective. Effective performance tasks, no matter their type, share several core characteristics. I outline these characteristics next.

Task elicits evidence aligned with theory of cognition about target of assessment. From the perspective of experts in educational assessment, foremost among the characteristics shared by effective performance tasks is that the task serves well in eliciting observations aligned with the theory of cognition about the construct or constructs targeted by the assessment (Pellegrino et al., 2001). For the assessment designer employing the Constructing Measures approach to the design and development of an assessment, this means that to have a chance at being effective, a performance task must be well-aligned well with the construct map (or maps) articulating the hypotheses of how those being assessed develop competency in the targeted construct (or constructs) (Wilson, 2005).

As a signal that this alignment exists, the instructions, directions, and details used in presenting a performance task to assessment takers—directions that intend to prompt the elicitation of construct-relevant evidence—will often include references to terms and actions featured in the relevant scoring guides used to interpret responses to the task (Wilson, 2005, p. 47). Since these scoring guides are aligned with the construct maps, the use of these terms and actions reflects alignment between the performance task and the construct map.

While this sort of alignment is considered a necessary condition for eliciting construct-relevant evidence, it does not guarantee that actual evidence elicited by a performance task will align with the hypotheses articulated in the construct map(s), nor, if it does, do so evenly across the population of assessment takers. Certain kinds of analyses of responses may reveal, for example, that for some sub-populations who took

the assessment, a performance task worked well to elicit construct-relevant evidence, but that for other sub-populations the same performance task did not successfully elicit construct-relevant evidence. Item-response modeling has numerous established methods for conducting such, and other, critical analyses for discerning and expressing the ability of a performance task to function effectively in such regards. These types of item analyses (e.g., to identify and quantify differential item functioning, if present) are essential for the arguments in support of an assessment's reliability and validity.

Focuses on an important aspect of construct being assessed, target of learning. This characteristic is related to the prior one, that good performance tasks will elicit evidence aligned with the theory of cognition for the target of assessment. A good performance tasks needs to focus on an important aspect of the construct being assessed. Data about other, less important, aspects of the construct may be elicited in the course of assessment takers engaging in performing the task. The main target of the task, however, when it is a "good performance task," needs to focus on an important aspect of the construct being assessed.

In K-12 contexts, these constructs are the important, overarching learning targets, or essential knowledge or "enduring understandings" (Wiggins & McTighe, p. 10) that are the goal of the learning activities. In the context of the P-P-P Assessment, these are the planning, enacting, and reflecting facets of the posing, pausing, and probing constructs. This is the rationale for the three performance tasks in the P-P-P Assessment's items design. Each of the three performance tasks focuses on a different facet of the three facets of teacher practice hypothesized in this study. The first performance task targets the

planning facet, the second targets the enactment facet, and the third targets the reflection facet.

Requires complex, authentic, realistic performances in context. A good performance task engenders complex and authentic performances of those being assessed. This contrasts with a task that requires those being assessed to apply specific and concrete skills in more de-contextualized and inauthentic settings, situations, or scenarios. In the context of an assessment designed to assess teachers' probing, for example, an example of the latter (i.e., a more de-contextualized and inauthentic setting for the task) might be a performance task that asks teachers to generate several probes that could be used to check for the presence of a particular misconception.

Asking teachers to generate probes in the absence of students may support teachers' learning and may be useful for learning about aspects of teachers' probing expertise with regard to a particular misconception, but it is not asking for a genuine teaching performance in context. To analogize with a sports-learning example: such a task would be akin to asking learners to dribble a basketball versus asking learners to play a basketball game. The former elicits application of a specific and discrete skill—dribbling—but differs from the latter in which dribbling is one of many applied skills combined in a complex and fast-changing context toward a goal, or perhaps toward several—even several multi-faceted—goals at once. Good performance tasks present realistic conditions for those being assessed to confront and navigate.

Directly meaningful to those being assessed. Good, "authentic" performance tasks are "directly meaningful" to those being assessed, as compared to being only "indirectly

meaningful" (Brookhart & Nitko, p. 261). Value judgments necessarily play a role in determining what is considered "directly meaningful." Questions arise: meaningful for whom? How meaningful is a performance task for respondents whose performances suggest confoundment or frustration with the task?

Answers to such questions have implications for the items design of an assessment, especially a performance-based assessment such as the P-P-P Assessment. For example, in the case of assessment takers who perform poorly on a performance task, it is recognized that there are known negative effects on motivation (Brookhart and Nitko, 2015). Though this statement both simplifies the interplay of contextual factors that likely influence this effect on motivation and leaves *motivation* undefined, it does raise an issue that becomes of increasing concern to many performance task-based assessments in a certain context, such as when large numbers of assessment takers engage the assessment. What if large numbers of teachers who would perform poorly with the performance tasks of the P-P-P Assessment were to complete them? To what extent might benefits of the use of the assessment outweigh unintended consequences?

This is one reason why employing the building blocks approach to assessment development, an item response modeling approach, is valuable. A powerful affordance of item response modeling is that it is possible to predict the properties of an assessment from the properties of the items of which it is composed (Pellegrino, Chudowsky, & Glaser, 2001). Having access to these kinds of predictions could facilitate the decision making processes of stakeholders involved in deciding who takes assessments such as the P-P-P Assessment.

The concept of a performance task being "directly meaningful" to those taking the assessment" has an inherently complex relationship with the items design of the assessment. The number of performance tasks featured in an assessment, for example, figures into this relationship. As previously mentioned, an items design needs to reflect the recognition that despite an assessment designer's best efforts, a single task alone cannot provide the amount and kinds of evidence needed to establish sufficient levels of reliability and validity for an assessment. An items design should also reflect the recognition that the potential for a single performance task to be "directly meaningful" to all assessment takers has limits too. This is another reason a thoughtful items design, one that may include strategic use of more than one performance task, is necessary.

Consistent with modern learning theory. Constructivist learning theory emphasizes that learners should use previous knowledge to build new knowledge structures, engage in participatory practices that support exploration and the embedding of knowledge in social contexts, and construct meaning for themselves (Hableton & Murphy, 1992; Rudner & Boston, 1994). A good performance task will be designed so that it expects and supports those being assessed in doing so. In its structure, such a performance task will take care to invite activation of assessment takers' prior knowledge and will prompt exploration and inquiry.

A good performance task will reflect the expectations that the knowledge and skills it aims to assess to be embedded by the person being assessed into a social context that is the assessment taker's present context, or close to it. In this way, and consonant with modern learning and assessment theories, good performance tasks reflect a situative, or

sociocultural, perspective (Pellegrino et al., 2001) in addition to embodying a cognitive perspective on learning. This perspective proposes that assessments are in part a measure of the degree to which one can participate in a form of practice (Pellegrino et al., 2001, p. 63).

For example, a good performance task within an assessment designed to assess teachers' proficiency in probing student thinking would not require teachers to instruct a class full of students previously unknown to them, for two reasons. Firstly, enactment of skillful probing requires that teachers know students' habitual reactions to direct probing versus indirect probing and to "public" versus "private" probing and enacting probing that uses this knowledge. Such a performance task would not be well-aligned with the construct map for probing. This is since teachers being "assessed" with such a task inherently would not have access to students' habitual reactions to different kinds of probing (e.g., direct or indirect probing) in different contexts (e.g., contexts of "public" probing versus "private" probing). Therefore, by definition, the teachers "being assessed" would not have opportunities to demonstrate an important aspect of probing articulated in the probing construct map. Such a performance task/item would fail to generate observations needed for assessing the level of "possession" of the construct in the individual being assessed—before the observation process has even begun.

Secondly, this context would be too far from a teacher's present context, in which the teacher "knows" her students.[2] A well-designed performance task gives those being

[2] A common exception, of course, is the beginning of the school year, when typically nearly all students are new and unfamiliar to a teacher. This exception points out that what defines a "good" performance task is related to factors such as when/how it is presented to and completed by those being assessed. The context

assessed the opportunity to apply their learning to a new situation (and not just repeat or regurgitate it, for example. At the same time, the definition of "new situation" needs to make sense for the construct or constructs targeted by the task and the assessment in which the performance task resides. A performance task that expects teachers to instruct students previously unknown to them is not a sound definition of applying their learning to a new situation.

Variety of response modes suited to purposes. Response mode refers to how those being assessed will communicate their knowledge and skill, such as by speaking, writing, or by performing physical actions other than speaking or writing (body-kinesthetic response mode). A performance task intended to generate observations about the writing skills of those being assessed will require writing. Depending on the purposes of the assessment in which this particular performance task is embedded, for such a task, no other response mode, such as oral report as part of the task, may be warranted. Performance tasks with other purposes, however, should include response modes besides having respondents write, lest the assessment run the risk of assessing takers' abilities to write as much or more than it assesses their possession of the theorized construct. Including a variety of response modes can guard against construct irrelevance, a validity threat.

In a good performance task, the variety of response modes prompted over the expected course of task completion is well-suited to the purposes of that particular

---

in which a performance task is given to those being assessed affects the effectiveness of its functionality. Protocols for the administration of items in an assessment need to reflect the assessment designer's having taken this into account in order for an assessment to function well in meeting its purposes.

performance task. The variety—or lack of variety—of response modes in a particular performance task will be sensibly related to the overall items design of the assessment. Performance tasks that rely on a single response mode run the risk of being biased against populations with, for example, learning disabilities or with cultural differences that are not aligned with the required response mode.

Require integration. Quality performance tasks require integration of knowledge, skills, and abilities (Hambleton & Murphy, 1992; Rudner & Boston, 1994). Demonstrating integration of knowledge, skills, and abilities is more challenging than demonstrating discrete knowledge and skills. By several general taxonomies of knowledge, skills, and abilities (e.g., Bloom's, Webb's) integrating is taxonomically higher-order. Good performance tasks will aim to assess such taxonomically higher-order cognitive processes. Such performance tasks are essential to the construction of an items design that, when instantiated into an assessment and taken by those being assessed, will function to generate the number and kinds of observations needed to provide sufficient evidence to support inferences about a complex, multidimensional phenomenon. Teacher practice in the domain of formative assessment is hypothesized to be a complex, multidimensional phenomenon. It bears repeating, however, that no single performance task, no matter how "good," authentic, complex, "directly meaningful," or realistic, even when the task "requires integration" can, by itself, serve to achieve minimum standards of reliability and validity for an assessment (Wilson, 2005).

Fair. Good performance tasks strive to be fair across the entire intended population of those being assessed. Some performances tasks may expect or require the use of

resources not provided in the description of performance task nor in its presentation to those being assessed. In such cases, performance task designers need to consider: how fair is the task? Will those being assessed have fair and equal access to the expected resources?

Equal access to resources is not equivalent to fair access. In the case of the P-P-P Assessment, the subjects being assessed were teaching with different curricula. While all the teachers had teaching guides for their curriculum (i. e., "equal access" to an expected resource), not all teaching guides may have been equally as focused on—or skilled at—identifying probing questions teachers could ask to uncover students' misconceptions about a topic. This has implications for an assessment that has as one of its purposes the evaluation of teachers' proficiency in planning for, enacting, and reflecting on probing.

Open. Good performance tasks do not yield, nor aim to yield, a single, correct answer. They are open (Brookhart & Nitko, 2015) in nature. Respondents may complete the task using a variety of approaches and strategies. The nature of evidence elicited by such open-response items tends to be multi-faceted and can play a strong role in helping to meet the demands of providing sufficient evidence to support inferences about complex, multidimensional phenomena.

The next section explains why performance tasks are used when an assessment's aim is evaluating teacher practice. It also presents the features of well-designed performance-assessments for teachers that are supported by research.

Why performance assessments are used for evaluating teacher practice. Many of the advantages of using performance assessments and performance tasks with K-12

students apply equally well when the individuals who are taking the assessment and doing the performance tasks are not K-12 students, but teachers. One of these advantages is that performance tasks clarify the meaning of complex learning objectives for those engaged in the task and for those who are witnessing, or who are directly (or indirectly) involved in the task with the individual being assessed. In the case of K-12 students, when performance tasks are shared with the parents of students, the learning goals become clear to parents (as well as students) "through actual example" (Brookhart & Nitko, 2015, p. 269). This is an important strength of performance tasks.

The parallel to this affordance in the context of teachers-as-assessment takers suggests that as teachers engage a performance task designed to elicit important evidence about a quality or qualities of their teaching practice, they gain clarity about the worthy goals, practices, and constructs that the performance task is playing a role in assessing. It also means that others, such as colleagues, instructional coaches, or principals, who may learn about the performance task from the teachers, may also gain clarity about the construct, or "learning goal" being assessed "through actual example" (Brookhart & Nitko, 2015, p. 269). When this latter instance occurs, and educational professionals beyond the individual teacher completing the performance task gain clarity about a construct important to quality teaching, it can be argued that this advantage of performance assessments has served to positively influence the professional context in which the teacher works.

Important characteristics of performance-based assessments of teacher practice. According to performance assessment and teacher evaluation expert, Linda Darling-

Hammond, lead editor of *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning* (2014) and author of *Getting Teacher Evaluation Right: What Really Matters for Effectiveness and Improvement* (2013), researchers have found that well-designed performance-based assessments of teacher practice:

> 1. **Capture teaching in action** by looking at classroom practice in terms of what both teachers and students are doing to achieve particular learning goals;
> 2. **Observe and assess aspects of teaching related to teachers' effectiveness**, such as activating and building on students' prior knowledge, creating appropriate scaffolds to support the steps of the learning process, and creating opportunities for students to apply their knowledge, receive feedback, and revise their work;
> 3. **Examine teachers' intentions and strategies** for meeting the needs of particular students and the demands of the subject matter being taught;
> 4. **Look at teaching in relation to student learning** by evaluating student work that results from teaching, plus teachers' feedback and support that further improves student work; and
> 5. **Use rubrics that vividly describe performance standards** at different levels of expertise to evaluate teachers' practice strategies, and outcomes (Darling-Hammond & Wei, 2009; Pecheone & Chung 2006). (p. 26) [bolded in original]

Darling-Hammond recommends that performance-based assessments of teacher practice share additional characteristics as well. Research on such assessments sharing this additional set of characteristics (listed next), assessments designed for and taken by wide populations of beginning and experienced teachers, has found that when these kinds of assessments are "used to guide teaching and provide teachers with feedback... teachers are able to improve their skills" (Darling-Hammond, 2013 p. 27). This "wide population" refers to teachers who took one of three performance assessments: (a) California's PACT, or Performance Assessment for California Teachers, for preservice, preliminary credential-seeking teachers; (b) Connecticut's BEST, or Beginning Educator Support and

Training, for second-year or third-year teachers seeking licensure; or (c) the National Board for Professional Teaching Standards' (NBPTS) assessment for experienced teachers to earn a national certification that signifies achievement of accomplished teaching.

All of these performance assessments of teaching practice are portfolios. They also all share the requirement that while completing the portfolios, teachers being assessed (a) collect evidence of actual instruction through video recordings, curriculum plans, and samples of student work and learning and (b) write commentaries explaining "the basis for their decisions about what and how they taught in light of their curriculum goals and student needs, and how they assessed learning and gave feedback to individual students" (p. 27). Significantly, as assessments that "both *document* and help teachers *develop* greater effectiveness" (p. 27) [emphasis in original], they are an instantiation of a type of assessment *as* learning, in counterpoint to assessment *of* learning.

Teachers taking these kinds of assessments report their engagement supports their professional learning and catalyzes changes in their practice (Athaneses, 1994). Observational studies have also documented that these changes in teacher practice do indeed occur (Lustick & Sykes, 2006; Sato, Wei & Darling-Hammond, 2008). This occurs for not just experienced teachers participating in such performance assessments, but also for beginning teachers. Furthermore, studies of teachers who engaged Connecticut's BEST and California's PACT have been found to also help beginning teachers improve their practice in ways that continue after the assessment experience has ended (Chung, 2008; Darling-Hammond, Newton & Wei, 2013).

In addition, another important aspect of such performance assessments is linked to a strength of performance tasks I have discussed: performance tasks help assessment takers *and those connected to them*—such as parents when K-12 students are doing the task; or colleagues, instructional coaches, and principals when teachers are the ones doing the task—gain clarity about important, worthy learning goals. The benefits of educational stakeholders gaining clarity—and possibly agreement—about good or effective teaching practice in a field that, as Grossman, McDonald (2008) and others have argued "suffer[s] the consequences" of "still lack[ing] a framework for teaching with well-defined common terms for describing and analyzing teaching" (p. 186), should not be underestimated. Darling-Hammond asserts that participation in the kinds of assessments that share the set of characteristics outlined "supports learning both for teachers who are being evaluated and for educators who are trained to serve as evaluators" (p. 27). Moreover, well-designed performance-based assessments with these characteristics that are used to guide teaching and provide teachers with feedback "create a common language and set of understandings about good teaching for the field as a whole" (p. 27).

The affordances of well-designed performance tasks make them a good open item format type to use within the items design of an assessment that is intended for teachers. The next section presents the timeline, a sample of the task analysis, and the table of specifications for the P-P-P Assessment.

Timeline, Task Analysis and Table of Specifications

The items design of the P-P-P Assessment sought to balance collection of evidence of teacher practice of the three dimensions of formative assessment hypothesized in this

study—pausing, pausing, and probing—from each phase of the cycle of inquiry: planning, enacting, and reflecting. This plan for data collection reflects that this study posits planning, enacting, and reflecting to be three distinct and critical facets of teacher practice of FA. To convey significant information about the items design of the P-P-P Assessment, this section presents (a) the timeline experienced by assessment takers, (b) results of the task analysis of two items that anchored collection of planning- and reflecting-related evidence: the common lesson planning template and the video-stimulated recall (VSR) protocol, and (c) the table of specifications for the P-P-P Assessment.

Timeline. Data were collected for each subject over 4-6 weeks from December 2016-March 2017. Teachers' individual participation in the P-P-P Assessment was staggered. Table 9 presents the items from the P-P-P Assessment in the order they were presented to respondents.

Table 9: *Timeline of the P-P-P Assessment Items*

Week	Phase		
	Planning	Enacting	Reflecting
1	Interview (SSIP-1) EAs: a) FA concept mapping b) FA moves wheel [55 items]		
2	Pre-Planning & Enactment Survey [48 items]		
3	PT 1: Plan lesson, use "FA Moves LP Tool"		
4		PT 2: Enact Lesson	
5			PT 3: Reflect on Lesson (SSIP-2) EAs: a) FA concept mapping b) FA moves wheel c) VSR [11 items]
6			Exit Survey [10 items]

Note. SSIP = semi-structured interview protocol; EA = embedded assessment; PT = performance task; VSR = video-stimulated recall protocol.

Teachers participated in two in-person semi-structured interviews lasting approximately 90 minutes each: one during the intake/planning session and one during the session dedicated to reflection. The first planning-oriented interview is referred to as SSIP-1, or semi-structured interview protocol number one. The latter reflection-oriented interview is referred to as SSIP-2, or semi-structured interview protocol number two.

Each interview included embedded assessments, or EAs, also shown in Table 9. Two of the EAs were presented in SSIP-1 and presented again to teachers in SSIP-2 so they could add, amend, or revise the responses they gave during SSIP-1. These items sought to elicit evidence of new understanding and perhaps evolving conceptualizations of FA practice experienced by participants. As Table 9 shows, these two EAs were the "FA concept mapping" item and the "FA moves wheel" item. The third EA in Table 9 is the video-stimulated recall (VSR) protocol. I describe the video-stimulated recall embedded assessment after presenting the task analysis of the lesson planning template.

Task analysis of lesson planning template and video-stimulated recall protocol. This next section describes two of the items within the P-P-P Assessment: a lesson planning template and the series of prompts used to stimulate reflection on two video clips from the respondent's enacted lesson. The first item was presented to respondents during the planning phase, the latter during the reflection phase. This was in accordance with the aim of the P-P-P Assessment's items design to collect evidence from all three phases of the cycle of inquiry (planning, enacting, reflecting).

This section also presents a summary of the task analysis of these two items. They serve as examples of the task analysis conducted on all the items/embedded assessments of the P-P-P Assessment. Task analysis is conducted to garner evidence about an assessment to use in making an argument for validity based on test content (AERA, APA, NCME, 2014, p. 14).

Task to use a specific lesson planning template: "FA Moves Lesson Planning Tool." For the first of three performance tasks in the P-P-P Assessment, teachers were

asked to plan a lesson using a lesson planning template introduced to them during the intake/planning interview session (SSIP-1). Figure 10 presents Leila's response to the "FA Moves Lesson Planning Tool."

LESSON OVERVIEW	
UNIT TITLE: Proportional Relationship	**GRADE LEVEL:** 7th
LESSON TITLE: Proportional Relationships with Tables and Graphs	**DURATION:** 100 min.
SUMMARY	Students will create tables, graph proportional relationships, and identify proportional relationships in them. They will begin to investigate other situations in which the relationship of two quantities is proportional. They will focus on the graphs of these situations, learning to recognize proportional relationships from a graph.
BACKGROUND	Students have learned from previous lessons that two quantities are related proportionally if the ratios of these quantities are the equivalent in any given time. Students also recognize proportional relationship of two quantities from a graph when it shows a straight line that goes through the origin. Students learn that proportional relationship is multiplicative and is not additive.
MATH STANDARDS	**Common Core STANDARDS:** 7.RP.2a. Decide whether two quantities are in a proportional relationship, e.g., by testing for equivalent ratios in a table or graphing on a coordinate plane and observing whether the graph is a straight line through the origin. 7.RP.2d. Explain what a point *(x, y)* on the graph of a proportional relationship means in terms of the situation, with special attention to the points (0, 0) and (1, *r)* where *r* is the unit rate. **ELD Standards:** Expressing information and ideas in formal oral presentations on academic topics SL7.4-6; L7.1, 3
STUDENT LEARNING OBJECTIVE(S)	SWBAT verbally explain how the relationship of two quantities is proportional, and show the proportional relationship of two quantities by plotting a linear graph that goes through the origin using data from a given table.
ACADEMIC	**LANGUAGE DEMANDS:** Instructional Strategies: Reciprocal

LANGUAGE	Teaching (Individuals articulate understanding) In pairs, Person A pretends that Person B was absent and explains a concept. Switch roles and continue. **NEW VOCABULARY:**
STRATEGIES/TOOLS TO MEET ALL STUDENT NEEDS	**LANGUAGE DEMANDS/ELL:** CPM's word problems are lengthy and wordy. We will use close reading strategy to make sense of the word problems. CPM cooperative learning – each student plays a CPM role (resources manager, reporter/recorder, task manager, and facilitator). **ADVANCED:** **LACKING PREREQUISITES:** Small group differentiated instructions
ASSESSMENT	In your Learning Log, explain what a proportional relationship is and how you can see it on a graph and in a table. Include diagrams to illustrate your thinking and make an example of your own.
REFERENCES	CPM Teacher Notes
TEACHER PREPARATION	
MATERIALS	CPM textbook. Lesson 4.2.2 Resource Page
ADVANCE PREPARATION	Questions: 1. "What does it mean for a relationship to be proportional?" 2. "What is the cost to you if you do not buy anything in a store?" 3. "Where on the graph can we show the cost for buying nothing?"

LESSON IMPLEMENTATION		
Components	**Activities**	**Assessment strategies and uses**
Engagement	I show students two graphs without x and y labels and let students come up with situations that can fit the description of these graphs.	Think-pair-share

Instructions	Reciprocal Teaching: "What does it mean for a relationship to be proportional?" Use Hot Potato strategy: problem 4-34 Teamwork: problem 4-35 a, b, c Class discussion: problem 4-35 d, e, f Cornell Note for 4-35f: The point (1, y) on the graph describes the cost for one pound of cheese. This is called the unit rate. "Unit" means 1.	I circulate the room to hear what the students are saying. I go to the table groups where the focal students are seated. CFU As I circulate the room, look for graphs that do not have straight line that passes through the origin. Ask these questions if I found graphs that missing data point (0,0). "What is the cost to you if you do not buy anything in a store?" "Where on the graph can we show the cost for buying nothing?"
	Group Discussions: problem 4-36 Cornell notes: When quantities are listed in a table, the relationship of two quantities is proportional only if the ratios of these quantities are equivalent, and it makes sense when both quantities are zero. When a graph is a straight line that goes through the origin, the relationship of the two quantities is proportional. Teamwork, class discussion: problem 4-37 Spread out: problem 4-38 a, b, c, d Each table group will do a part of the problem, and then share out.	Share-out Select a group randomly by picking sticks. I circulate the room to hear what the students are saying. I go to the table groups where the focal students are seated. The reporter from each table group will share out what his or her team has found.
Closure	In your Learning Log, explain what a proportional relationship is	

	and how you can see it on a graph and in a table. Include diagrams to illustrate your thinking and make an example of your own. Title this entry "Proportional Relationships in Graphs and Tables" and label it with today's date.	
Small group differentiated instructions	While students are working on writing the learning log, I will work with a small group of students to help them recognize whether the relationship of two quantities is proportional or not.	I teach students the way to write ratios, find simplest form, and plot graph. I use sentence frames to help students articulate their understanding of the concept of proportional relationship

Figure 10. Leila's response to the lesson planning template.

The "FA Moves Lesson Planning Tool" item sought to generate evidence of respondents' planning for posing, pausing, and probing. It targeted all five of the levels hypothesized in the respective construct maps. Respondents were given paper and electronic document versions of the Tool during SSIP-1. They were prompted, "Please use this 'FA Moves Lesson Planning Tool' as you plan. I will be collecting a copy of what you come up with."

Respondents scheduled the enactment—and video recording—of the planned lesson to occur within one to three weeks of receiving the *Tool*. No other time constraints were placed on the task. If a respondent asked about receiving feedback on a response, then minimal feedback was given. One of six study subjects asked for feedback. In reply, and one day after the respondent shared the lesson plan for feedback, the respondent received two positive comments and one question by email ten days before the scheduled

enactment of the lesson. The promptness and timing of the reply was intended to allow the respondent sufficient time to consider and act upon the feedback before the scheduled enactment of the lesson plan.

Table 10 presents the task analysis of the FA Moves Lesson Planning Tool item. Conducting task analysis provides evidence for validity based on test content.

Table 10: *Task Analysis of FA Moves Lesson Planning Tool Item*

Task Demands	Cognitive Demands	Openness	Complexity
Read *Tool*; interpret diagram (template); determine uses for text boxes by interpreting categories in *Tool* (e.g., *language demands, components*); identify examples of these categories in intended lesson; express plans in writing within given categories	Regulate relationship between extant lesson plan (such as a lesson provided by curriculum/Teacher's Guide) and the expectations of a lesson plan expressed in the *Tool*; selectively transfer content of extant plan(s) to *Tool*; determine if additional lesson planning content needs to be added	Moderate level of constraint on the problem space, i.e., instructed to identify "student learning objectives(s)", "new vocabulary", etc. in text boxes in "Lesson Overview" section; expected to use text boxes with three columns (Components, Activities, and Assessment strategies and uses) provided in "Lesson Implementation" section; option to use paper or electronic version of *Tool*.	Light reading load; moderate visual load (three sections in *Tool,* each with multiple text boxes); moderate (semantic) language load (the *uses* of the words in the template—not the words themselves—may or may not be new to respondent, e.g., *language demands, components, assessment uses*)

Protocol to stimulate reflection on posing, pausing, and probing moves. For the third performance task of the P-P-P Assessment, teachers were prompted to reflect on two (or perhaps three) 3-5 minute long video clips of their enacted lesson while these clips were played back during the second and final "reflection interview" session. The

video clips respondents reviewed during the interview session were chosen to feature student-teacher interactions during both whole class and small group configurations. Generally, this meant that respondents were asked to reflect on one whole class configuration-focused video clip and one small group configuration-focused clip.

Additionally, all teachers were invited ahead of the session to choose a clip for reflection during the interview. If a respondent did choose a clip, the VSR protocol portion of the interview opened with re-play—and teacher "unpacking" —of the clip the teacher had chosen. Two-thirds of respondents chose a clip to reflect on during the reflection interview.

Sixty minutes of the 90-minute "reflection interview" session were budgeted for the video-stimulated recall (VSR) protocol. Figure 11 presents the protocol. The structure of the protocol reflects the option for teachers to begin by reflecting on a video clip of their choosing. Note that the introductory phrases used before each video clip playback differ slightly, while the questions and statements as options to use during video playback—the "Pose/Launch" questions and "Probes" within the shaded box—are the same. The five "pose/launch" questions were to orient the assessment taker to the task generally and to elicit initial responses. The five probes were structured to elicit elaboration and built on one another. Together they sought to elicit contextual information that was helpful to assessing the recorded performance and other evidence that the respondent might use in the probes that invited respondents' thoughts about possible next steps.

—By email I invited you to choose a clip (3-5 minutes long) that you would like to discuss in our session today.

—Which clip have you chosen?*

[Analytic note*: If the subject has chosen the chosen the "same" clip as one of the clips the researcher has chosen, note how the starting and ending points compare.]

—When does the clip begin? [Note time marker. Go to that point on video.]

—What is the ending point of your clip? [Note.]

—Before we watch the clip together, please tell me, why did you choose this clip?

—What is it about the posing, pausing, and/or probing in this clip that you'd like to talk about?

Before pressing play at the start of the clip the teacher has chosen, say:

As we watch, the goal is for you to reflect and "unpack" the clip. Stop the recording at any place. I am really interested to hear your thinking and decision-making related to the posing, pausing, and probing in this clip. Feel free to talk about **what you did** and would have **wanted to do, if anything,** at a particular moment in the clip.

Teacher presses "play." Teacher presses "pause" and reflects out loud.

[use 5 poses and probes in the box below as called for]

For the clips the researcher has chosen:

Before pressing play for "Researcher Clip 1 (Whole Class)", say:

As we watch this clip, the goal is the same as before: for you to reflect and "unpack" the clip. Please talk about your thinking and decision-making related to posing, pausing, and probing. What were you thinking? Intending? What were you anticipating or surprised by?

As we go through these clips, feel free to talk about **what you anticipated**, **what you did** in the moment, and how, if at all, **you wanted to do something different**.

—This clip comes from the part of the lesson where…

—It begins with…

—Press "pause" at any time.

Researcher presses "play." Teacher/researcher presses "pause" and teacher reflects out loud.

For "Researcher Clip 2 (Small Group)", say:

Though this is a different clip, the goal is the same as before: for you to reflect and "unpack" the clip related to posing, pausing, and probing. Please talk about your thinking and decision-making related to posing, pausing, and probing. What were you thinking? Intending? What were you anticipating or surprised by?

As we go through these clips, feel free to talk about **what you anticipated**, **what you did** in the moment, and how, if at all, **you wanted to do something different**.

—This clip comes from the part of the lesson where…

—It begins with…

—Press "pause" at any time.

Researcher presses "play." Teacher/researcher presses "pause" and teacher reflects

out loud.

Pose/Launch:

1. What are you noticing about posing-pausing-probing in this clip? Does anything stick out?
2. Looking at this clip, do you having any thoughts NOW about planning this clip/segment/episode?
3. What do you wish you been able to do in this moment as you enacted the lesson segment/episode? Please explain. How does this relate to posing-pausing-probing or the FA moves generally?
4. If you were to fast forward and teach this [e.g., posing or pausing or probing routine] again, what you want to do?
5. Considering this video overall, do you have any other thoughts or reflections?

Probes:

- Say more… [elaboration]
- In terms of posing/pausing/probing, what are we not seeing that is important to consider? [what's invisible?]
- Can you offer another explanation? [other explanations/possibilities]
- Tell me about your purposes for doing ____. [purposes]
- In terms of possible things to try next re: posing, pausing, or probing, what might you suggest? [contingency]

Figure 11. Video-stimulated recall protocol.

Using this protocol, four of six total respondents reflected on two clips using this protocol during the reflection interview. Two respondents were able to reflect on three video clips using the protocol. In all but one case the respondents reflected on clips that featured small group and whole class configurations. Only one respondent had clips of only whole class instructional configurations to reflect on. This respondent's enacted lesson included no student-teacher interactions within small group configurations.

Table 11 presents the task analysis of the VSR protocol. A noteworthy aspect of this protocol is that it demands no reading or writing from the assessment taker. A constraint of the protocol relates to the timing of its presentation to assessment takers (Calderhead,

1981). Presentation of the video clips to assessment takers sought to be close enough in temporal proximity to the lesson enactment that thoughts that had occurred to the respondent during the lesson enactment remained accessible. In the present study, no response processes-focused items that targeted the teachers' experiences of this aspect of the P-P-P Assessment's items design were included. Future studies that employ the VSR protocol should seek to elicit such evidence from assessment takers.

Table 11*: Task Analysis of Video-stimulated Recall Protocol*

Task Demands	Cognitive Demands	Openness	Complexity
Watch video; recognize FA moves enacted or attempted; recall thoughts and/or decisions that during video recording; initiate stops and starts of video playback; provide contextual information to researcher; speak to purposes/intentions behind actions	Process oral prompts; hold focus on prompts while watching/reviewing video; distinguish between present thoughts/reflections and thoughts/decisions that occurred during enactment; explain own actions; hypothesize explanations for student behavior; generate possible "next steps"	Moderate level of constraint on the problem space, i.e., view through lens focused on posing, pausing, and probing and other FA moves	No reading load; high visual load, i.e., video clips

Table of specifications. Table 12 presents the table of specifications for P-P-P Assessment. The top four rows of Table 12 display the items that targeted the domain of FA practice in general and the three hypothesized facets of FA practice generally. If an item, for example, asked about planning in general, rather than ask specifically about planning for posing or planning for pausing, it was identified in this top quarter of Table

12. The rows of the bottom three-fourths of Table 12 identify the items that targeted each dimension of FA practice hypothesized in this study in general—posing, pausing, and probing—and each hypothesized facet within each of those dimensions (i.e., planning for posing, enacting posing, reflecting on posing, planning for pausing, enacting pausing, reflecting on pausing, planning for probing, enacting probing, reflecting on probing).

As Table 12 shows, the items design for the P-P-P Assessment included several item types. Items in the P-P-P Assessment included three types of fixed choice items (Likert-scaled, ranking, and frequency), four types of constructed response items (short answer, concept mapping, the "FA moves wheel" item, and oral responses to interview questions), and three performance tasks (one for each facet of practice hypothesized). The "FA moves wheel" item asked assessment takers to connect elements of their teaching practice to the FA moves as they were examining a graphic of Figure 3, the "FA moves wheel." See Figure 3 on page 23 in chapter 1 for the FA moves wheel, which defines the FA moves.

Table 12: *Table of Specifications for P-P-P Assessment*

To Assess	**Item Types**				
	Fixed Choice		Constructed Response		Performance Tasks
Domain	Likert	Ranking or frequency	Short answer	Concept map, FAMW, oral	(1) LP template, (2) enact lesson, (3) VSR
FA (general)			PreS: 1, 2, 3, 45	concept map in SSIP-1	PT1, PT2, PT3
Planning				SSIP-1: 11	PT1: LP template
Enactment					PT2
Reflection				concept map in SSIP-2; FAMW (SSIP-2)	PT3: VSR
Dimensions					
Posing (general)	PreS: 5	PreS: 6, 7, 8, 10, 13	PreS: 4, 9, 11, 14, 15, 16, 45	FAMW (SSIP-1); SSIP-1: 3, 4	PT1, PT2, PT3
Planning for Posing		PreS: 12		SSIP-1: 2	PT1: LP template
Enactment					PT2
Reflection on Posing				FAMW (SSIP-2)	PT3: VSR
Pausing (general)	PreS: 33	PreS: 34, 35, 37, 38, 41	Pre: 32, 36, 39, 42, 43, 44, 45	FAMW (SSIP-1); SSIP-1: 6	PT1, PT2, PT3
Planning for Pausing		PreS: 40		SSIP-1: 5, 7	PT1: LP template
Enactment					PT2
Reflection on Pausing				FAMW (SSIP-2)	PT3: VSR
Probing (general)	PreS: 18	PreS: 19, 20, 22, 23, 25, 28	PreS: 17, 21, 24, 26, 29, 30, 31, 45	FAMW (SSIP-1); SSIP-1: 9	PT1, PT2, PT3
Planning for Probing		PreS: 27		SSIP-1: 1,10	PT1: LP template
Enactment					PT2
Reflection on Probing				FAMW (SSIP-2)	PT3: VSR

Note. FAMW = FA moves wheel; LP = Lesson plan; VSR = Video-stimulated recall; PreS = Pre-lesson enactment survey; PT= Performance Task; SSIP = Semi-structured Interview Protocol.

Each performance task was aligned with a phase in the cycle of inquiry. There was a performance task focused on planning, a performance tasked focused on enacting, and a performance task focused on reflecting.

By describing the characteristics of good performance tasks, and explaining the rationale for employing performance tasks when an aim is to assess teaching practice, this chapter has argued for the soundness of the items design for the P-P-P Assessment. The next chapter, chapter 4, is dedicated to the third building block in Wilson's Constructing Measures framework: outcome space. The development of the scoring guides that were used to assess the study subjects' responses to the three performance tasks that anchored the items design of the P-P-P Assessment will be described in chapter 4 (see appendix B for final versions of all three scoring guides).

Chapter 4: Outcome Space

This chapter (a) defines the concept *outcome space*, (b) contextualizes the outcome space design of the P-P-P Assessment in the field of K-12 education, and (c) describes the outcome space for the P-P-P Assessment and explains the rationale for its design. The first section presents the qualities of a sound and useful outcome space and three approaches commonly used to develop one. It introduces the necessary step an assessment designer must take in the development of an outcome space of relating the categories of the outcome space back to the generating construct map and discusses how and why this is done. It also explains that while the development of an outcome space involves the scoring of item-response categories, the concept of an outcome space is not wholly synonymous with creating or deciding on a scoring strategy and further explores the importance of this concept.

The second section of this chapter examines the corresponding concept of an outcome space in Wylie and Lyon's conceptualization of "Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice" for K-12 teachers, known as FARROP. When the P-P-P Assessment's outcome space is described in this chapter's final section, the differences between Wylie and Lyon's rubrics and the outcome space of the P-P-P Assessment become apparent. The ordered categories of the P-P-P Assessment's outcome space were related back to the generating construct maps during development of the outcome space; the rubrics in the FARROP scheme/conception were not. The chapter's conclusion explores the implications of these facts.

The third and final section of this chapter presents and explains early and final versions of scoring guides created for the P-P-P Assessment since three general scoring guides are central to the design of its outcome space. The drafting and revision of the P-P-P Assessment's outcome space aimed (a) to reflect qualities and procedures advocated by Wilson's "building blocks" approach to constructing measures and (b) to play an integral role in the generation of meaningful, "next steps" feedback to teachers, both teachers who took the P-P-P Assessment and teachers who one day might take future iterations of it. The contexts in which such future teachers will teach are expected to vary widely. The P-P-P Assessment's scoring guides will likely be applied to teacher practice of formative assessment outside the discipline of mathematics. Outcome space design decisions reflected this possible future application of the scoring guides. The feedback to the subjects of this study that the outcome space design of the P-P-P Assessment helped to generate is presented and explored in Chapter 5, "Profiles of Practice and Feedback."

Definition

An outcome space is a set of categories describing observations—responses to items or tasks—that serve to help in scoring—interpreting—them to be indicators of the targeted construct (Masters & Wilson, 1997; Wilson, 2005). Wilson (2005) characterizes a "sound and useful" outcome space as having categories that are "well defined, finite and exhaustive, ordered, context-specific, and research-based" (p. 62). The categories of an outcome space should be qualitatively distinct from one another.

Origins. Ference Marton (1981) introduced the term outcome space to describe a set of "outcome categories" for students' responses to a task developed using a particular

kind of analysis, known as phenomenography. Phenomenography, entailing detailed analysis of a range of student responses to standardized open-ended items, originated with Marton (1981, 1986, 1988). In Marton's use of the term outcome space, the phenomenographic analysis used to develop outcome categories—which, when taken all together comprise an outcome space—valued "discovering the qualitatively different ways in which students respond to a task" (Wilson, 2005, p. 63) (italics in original). Marton's work found that students' responses to a cognitive task invariably reflect a limited number of qualitatively different ways students think about a phenomenon, concept, or principle (Marton, 1988). The job of a "designer" of an outcome space—in Marton's conceptualization—is to discover this limited number of ways through phenomenographic analysis, categorize them, and present the categories in terms of some hierarchy.

Need for contextual understanding. The concept behind my use of the term here, however, and how Masters and Wilson (1997) define outcome space, is broader. One implication is that other strategies for coming up with and refining outcome categories—besides conducting phenomenographic analysis—may be used to develop an outcome space. Another implication is that use of the concept cannot be separated from Wilson's Constructing Measures framework (2005).

My purpose for using the term and concept relies on Wilson's definition, use, and explication of *outcome space* within his Constructing Measures framework or "building blocks" approach to designing assessments (2005). Therefore every aspect of an outcome space cannot be—nor should be—separated from its role and value in the process of

designing and developing an assessment from this particular item-response modeling approach. This suggests that understanding outcome space as a concept requires recognizing the role the properties of an outcome space (e.g., well defined, finite and exhaustive, context specific) can play in improving—or worsening—rater reliability and in providing evidence toward strengthening—or weakening—an assessment's or test's content validity argument. Rater reliability and the argument for validity from evidence based on test content (sometimes referred to as test content validity) are among the most significant indicators of the quality of an assessment; they are fundamental considerations in developing and evaluating tests according the the 2014 *Standards for Educational and Psychological Testing* (AERA, APA, NCME).

It also means that understanding *outcome space* requires knowing how an outcome space relates to the other aspects of the measurement framework, or "building blocks" in Wilson's Constructing Measures framework: construct maps, items design, and measurement model. The "building blocks" bear important relationships with one another. When the nature of these relationships is well understood, this understanding can be used to marshall evidence that supports the construction of a multi-faceted, solid, integrated argument for validity regarding an assessment's use.

Building blocks correspond to NRC's assessment triangle. This inter-relation of the building blocks to support a solid and integrated argument for validity corresponds to the National Resource Council's Committee on the Foundations of Assessment requirement that the three foundational elements that underlie all assessments "must be explicitly connected and designed as a coordinated whole" to support the inferences that

can be drawn from an assessment (Pellegrino, Chudowsky, & Glaser, 2001, p. 2 & p. 54). The NRC committee communicates these three "foundational elements" and their inter-relationship through its "assessment triangle."

The assessment triangle, with its vertices of cognition, observation, and interpretation, is useful in communicating the requirement that these three elements— (a) a model of ways in which students represent knowledge and develop competence in the subject domain, that is, a model of cognition, (b) tasks or situations that allow one to observe students' performance: observation, and (c) an interpretation method for drawing inferences from the performance evidence thus obtained: interpretation—must function "in synchrony...for an assessment to be effective" (p. 44). If these elements, which correspond to the construct maps, items design, and outcome space building blocks in Wilson's Constructing Measures framework, are not connected and designed as a coordinated whole, the NRC warns, "the meaningfulness of inferences drawn from the assessment will be compromised (p. 2 & 54).

Three common approaches to developing an outcome space. When employing the "building blocks" approach to devise an assessment, designers commonly use three general approaches to construct an outcome space: (a) phenomenography, (Marton, 1981; Uljens, 1993; Van Manen, 1990), (b) the Structure of Observed Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982; Van Rossum & Schenk, 1984), and (c) Guttman and Likert-item scales (Wilson, 2005). The lattermost approach is applicable to non-cognitive contexts and used for the creation of outcome spaces in attitude and behavior surveys. Choice of approach depends, among other concerns, on the

assessment's purposes and the assessment problems the designer is attempting to resolve. Using more than one approach to construct an outcome space may be helpful. Relying on only one approach, however, may be sufficient for some assessment development contexts.

Significance of "research-based categories." As mentioned, for an outcome space to be considered "sound and useful," the categories the outcome defines must be based on research. Taken together, the categories defined by an outcome space represent a model of cognition and learning in a domain. As was asserted and explained in Chapter Two, the Construct Maps chapter, the importance and centrality of this model of cognition to the assessment design process cannot be overstated. This centrality is in line with recommendations of a body of experts in educational assessment represented by the National Research Council's Committee on the Foundations of Assessment. In 2001 this committee proposed an approach to the design of educational assessments in Knowing What Students Know: The Science and Design of Educational Assessment that differed from most approaches being taken in the field at that time. "One of the main features that distinguishes the committee's proposed approach to assessment design from current approaches," the committee asserted as it contextualized the committee's proposal in the executive summary of its volume, "is the central role of a model of cognition and learning" (p. 6).

Wilson's requirements of an outcome space are aligned with the NRC committee's recommended approach. Wilson has argued, "a research-based model of cognition and learning should be the foundation for the definition of the construct, and hence also for

the design of the outcome space and the development of items" (2005, p. 66). The members of the NRC's Committee on the Foundations of Assessment have similarly argued, "A model of cognition and learning should serve as the cornerstone of the assessment design process." Moreover, these assessment experts put conditions on the model of cognition and learning and what it should reflect:

> This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain.…This model may be fine-grained and very elaborate or more coarsely grained, depending on the purpose of the assessment, but it should always be based on empirical studies of learners in a domain. Ideally, the model will always provide a developmental perspective, showing typical ways in which learners progress toward competence (pp. 3-6).

This suggests that when the model of cognition and learning represented by the categories of an outcome space is based on empirical studies of how learners progress toward competence in the domain or construct of interest—as Wilson's definition of *outcome space* requires the categories of an outcome space *must* be—the assessment designed with this model is better positioned to meet expectations for quality and process that educational assessment experts hold regarding assessments and their design.

Relating categories back to generating construct map. An essential step in creating an outcome space is relating the categories back to the responses side of the generating construct map (Wilson, 2005, p. 69). At one level, taking this step can be seen as the process of assigning numbers to the ordered levels of the outcome space—deciding on a scoring system for the item-response categories. But there is another level, a "deeper meaning" to this step that is critically important (Wilson, 2005, p. 69) for an assessment designer to understand. Seeing this step of relating the categories back to the responses

side of the generating construct map as "integral to definition of the categories" themselves is this "deeper meaning."

Significance of the relationship between outcome space and construct map. Recognizing this aspect of taking this step acknowledges the importance of the relationship between the outcome space and the construct map. In terms of the NRC committee's assessment triangle, recognizing this aspect pertains to understanding the necessity that the interpretation and cognition vertices, which correspond to Wilson's outcome space and construct map building blocks, sensibly relate to one another. They must, according the NRC's report, relate in ways that arguably align with the purposes of the assessment in order to support the meaningfulness of the inferences that can be drawn from it.

Categorization and scoring are not equivalent and why this matters. Part of understanding the relationship between an outcome space and its generating construct map is understanding that categorization and scoring are not the same—and knowing that they should work in concert with one another. This is critical to being able to use schematic knowledge of assessment design to work strategically to support the quality of the instrument/assessment being designed (Duckor, Draney, & Wilson, 2009).

Two additional reasons support how important it is for an assessment designer to distinguish categorization from scoring. First, being able to justify each step in the process of developing the assessment is essential to presenting a coherent argument for the assessment's design. Second, sometimes employing different scoring schemes yields insight into the latent construct. In this case, the categories would not change, but the

schemes for scoring them would differ. Knowing how and why particular scoring schemes applied to the data function better than others and, further, how the choice of outcome space is tied to both the items design and the nature of the information sought about the construct reflects advanced understanding of the creation and role of an outcome space in the design and development of assessments (Duckor, Draney, & Wilson, 2009).

How the Posing, Pausing, and Probing Scoring Guides Fit in the Field

This section aims to situate the posing, pausing, and probing scoring guides of the P-P-P Assessment within the field of K-12 education. Since future application of the P-P-P Assessment is expected to include instructional contexts across grade levels and subject areas—not just middle school mathematics—and since development of its outcome space took this into account, it is fitting to compare the P-P-P Assessment's outcome space to schemes in K-12 education that describe teacher practice across grade levels and subject areas. Therefore this review does not address well-known schemes for describing and evaluating teacher practice that focus only on mathematics instruction, such as the Inside the Classroom (ITC Observation and Analytic Protocol) (Horizon Research, 2000), the Learning Mathematics for Teaching: Quality of Mathematics Instructions (LMT-QMI) (Learning Mathematics for Teaching, 2006), the Instructional Quality Assessment (IQA) Mathematics Toolkit (Matsumura, Garnier, Slater, & Boston 2008; Boston & Wolf, 2006), and the Mathematical Quality of Instruction (MQI; Hill et al., 2008).

FARROP's "singular" focus. Chapter two introduced four of the most well-known schemes in K-12 education that describe teacher practice across subjects and grades that

include some aspects of the constructs posing, pausing, and probing as defined in this study: (a) Danielson's Framework for Teaching (FFT; 1996, 2007, 2011, 2013), (b) the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers' (CCSSO) "Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice" (FARROP) (Wylie & Lyon, 2013), (c) the Interstate New Teachers Assessment and Support Consortium's Learning Progressions for Teachers (LPfTs) 1.0 (CCSSO, 2013), and (d) the Teaching for Robust Understanding (TRU) Framework (Schoenfeld, 2014). Of these four, only FARROP purports to focus on the construct "formative assessment" and will be reviewed here.

Expectations for schemes not employing the Constructing Measures approach. It should be noted that none of these articulations of teacher practice were developed using the "building blocks," or Constructing Measures, approach. Therefore, it should not be expected that the process that each set of creators employed to establish/develop a scheme's set of ordered categories—in cases where information on this process is available—will reflect Wilson's recommendations for creating an outcome space. Likewise, one should not necessarily expect, then, that Wilson's requirements for a sound and useful outcome space have been met.

If all the requirements for a sound and useful outcome space are found to be met by a scheme, this would be a function of a different assessment design and development approach having successfully achieved what the "building blocks" approach—when employed skillfully—has reliably achieved in past cases of assessment construction (IEY,

Science Education for Public Understanding Program, 1995; LBC, Claesgens, Scalise, Draney, Wilson, & Stacey, 2002). Such would be a noteworthy finding regarding a scheme, given that the case has been already been made for Wilson's approach to constructing measures, which—when well executed—should satisfy the expectations for assessment quality as recommended by the National Research Council's Committee on the Foundations of Assessment (Pellegrino, Chudowsky, and Glaser, 2001) as well as position an assessment to be able to meet the field's standards for educational assessment published in the 2014 *Standards for Educational and Psychological Testing* (AERA, APA, NCME).

The question of situating the scoring guides from the P-P-P Assessment in the field becomes, then, a larger, but just as important question: how well do these other schemes satisfy these recognized and respected expectations, recommendations, and standards of educational assessment? The following description of FARROP's corresponding concept of an outcome space should serve to answer in part.

In describing FARROP's corresponding concept to an outcome space, I begin by putting aside, temporarily, the requirements of a sound and useful outcome space: that the categories comprising the outcome space be well defined, finite and exhaustive, ordered, context-specific and research-based. If, then, an outcome space can be said to be a set of ordered categories, then FARROP's outcome space is a set of ten rubrics where each rubric delineates four levels of practice: beginning, developing, progressing, and extending.

On FARROP's ten "dimensions." Each rubric is purported to describe a different "dimension" of formative assessment (Wylie & Lyon, 2013). Yet no validity evidence based on internal structure of the observation tools, rubrics, and protocols when used as intended has been published. Strictly speaking, these ten "dimensions" cannot accurately be called dimensions of formative assessment if appropriate analyses examining the degree to which "the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" have not been conducted (APA, AERA, NCME, 2014, p. 26). The specific types of analyses called for (e.g., factor or item-response theory analyses) depend on how the assessment will be used (APA, AERA, NCME, 2014). According to the 2014 Testing Standards, such examinations of validity evidence based on internal structure are one of five sources of evidence important to consider when constructing a coherent argument for validity of an assessment: (a) evidence based on test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) evidence for validity and consequences of testing.

In the case of FARROP, such examinations of internal structure would entail analyzing the extent to which the observation tools, rubrics, and protocols function in bearing out the presumptions of the framework: that there are ten dimensions of formative assessment and that the evaluator's observing a teacher teach, using the rubrics, and following the protocols will lead to *formative* evaluation of teacher practice regarding these ten dimensions, formative evaluation that supports a teacher's professional reflection on practice. Literature about FARROP clearly communicates that the rubrics

have not been developed for summative evaluations and "should not be used for that purpose without first studying their validity and reliability, creating a training and certification system for observers, and developing a process to monitor observer accuracy on an ongoing basis" (Wylie & Lyon, 2013, p. 4).

This begs the question: other than the "stakes" ostensibly being lower in "non-summative"—or formative—evaluation uses of FARROP compared to summative evaluation uses, what makes the observation tools', rubrics' and protocols' use for *formative* evaluation any sounder, psychometrically speaking, than its recommended *non-use* for summative evaluation purposes until studies, training, and validation have been carried out? Upon what evidence are the recommendations for using FARROP for formative evaluation based?

On FARROP's "model of cognition" and evidence base. Wiley and Lyon (2013) report that "The levels [in the rubrics] are referred to both by names and by numbers to indicate a progression of skills and abilities" (p. 13). It is not clear, however, that this articulation of a "progression of skills and abilities" is based on a model of cognition or studies of empirical evidence, as the National Research Council's Committee on the Foundations of Assessments recommends for all educational assessments (Pellegrino, Chudowsky, & Glaser, 2001). I could find no available evidence that identified and explained the model—or models—of cognition, or developmental progress, on which the rubrics were based. Nor could I find any available evidence demonstrating that empirical studies of teacher practice were used during the creation of the rubrics to support the

process of deciding upon the categorization and content of the categories in the 4-leveled rubrics.

Wylie and Lyon reported that 23 teachers reviewed the document, that several members of the FAST SCASS provided helpful feedback and suggestions, and that some of these shared "state materials" that informed the content of FARROP (Wylie & Lyon, 2013, p. 2). For additional information on the construct formative assessment, the FARROP document (Wylie & Lyon, 2013) directs readers to two prior FAST SCASS publications about formative assessment. One of these publications is, essentially, 13 vignettes of teacher practice "taken from teacher observations conducted in a variety of schools across the U.S." (Wylie, 2008, p. 3). The teacher observations that led to the articulation of these 13 vignettes appear to be the extent of empirical evidence of teacher practice informing the creation of the rubrics.

In the document that provides guidelines and resources for using the FARROP framework with teachers, Wylie and Lyon (2013) state "there are also a variety of texts on formative assessment that represent the key ideas in a way that is congruent with the FAST SCASS definition," footnoting Heritage's (2010) *Formative Assessment: Making it Happen in the Classroom* and Popham's (2008) *Transformative Assessment*. But reporting that key ideas in two books are "congruent" with the definition of formative assessment they relied upon—FAST SCASS's definition (McManus, 2008) —when they drafted and revised the rubrics *is not the same* as citing empirical studies that informed a model of cognition grounding an assessment's creation. It is not the same as citing empirical studies that informed the development of the set of ordered categories arrived

at amongst FARROP's ten rubrics, i.e., describing the research base of the research-based categories of an outcome space.

Variable clarification may be needed. Table 13 presents the rubric for "dimension" IV in FARROP, "Questioning Strategies That Elicit Evidence of Student Learning." Notice that as a construct "questioning strategies" seems to encompass aspects of posing, pausing, and probing. It may be that analyses examining internal structure of FARROP will support "questioning strategies" as a sound construct for the purposes for which FARROP was created. However, to an assessment designer who followed Wilson's advice to take one characteristic, or construct, at a time in order to see each as a construct map, "questioning strategies that elicit evidence of student learning" seems to conflate at least three constructs. Variable clarification, a requirement of designing and building an assessment—even an assessment intended for formative evaluation—that will work well seems called for in this case (Wilson, 2005, p. 38). To my knowledge, no evidence supporting the soundness of the "questioning strategies" construct/dimension of formative assessment as "a researched ordering of qualitatively different levels of performance focusing on one characteristic (Wilson, 2009, p. 3) has yet been published.

Table 13: *Rubric for Dimension IV, "Questioning Strategies That Elicit Evidence of Student Learning" in FARROP*

1-Beginning	2-Developing	3-Progressing	4-Extending
The teacher **asks** very few **questions** designed to assess student progress.	The teacher **asks** some **questions** at appropriate points to assess student progress.	The teacher **asks questions** at appropriate points to assess student progress.	Same as level 3
The teacher provides inadequate **wait time** and/or often answers own questions.	The teacher *inconsistently* provides adequate **wait time** to allow all students to engage with the questions. The teacher sometimes answers own questions.	The teacher provides appropriate **wait time** to allow all students to engage with the questions.	Same as level 3
The teacher uses questioning strategies that provide evidence from only a few students or the same students in the class.	The teacher *inconsistently* uses questioning strategies to collect evidence of learning from *more* students (e.g., whiteboards, exit tickets, etc.) but implementation may not be consistent or structured in a beneficial way.	The teacher uses effective questioning strategies to collect evidence of learning from *all* students in systematic ways (e.g., whiteboards, exit tickets, etc.)	Same as level 3
The evidence collected cannot be used to make meaningful inferences about the class's progress on intended learning outcome and to adapt/continue instruction.	The teacher misses multiple *critical* opportunities to make inferences about student progress and/or adapt/continue instruction accordingly.	The teacher occasionally misses *critical* opportunities to make inferences about student progress and adapt/continue instruction accordingly.	The teacher effectively uses student responses, **probing** for more information as necessary, to make inferences about student progress and adjust/continue instruction accordingly.

Note. Intersections with posing, pausing, and probing are bolded.

Both the P-P-P Assessment and the FARROP are intended for application to instructional practice of a wide range of grades and subjects. A critical difference between the P-P-P Assessment's outcome space, however, and the FARROP corresponding concept to an outcome space (the categories and content of its ten rubrics) is the P-P-P Assessment's integration of teacher planning, enactment and reflection into its outcome space. For the P-P-P Assessment these three "facets of practice," as this study conceptualizes and calls them, reside within the construct maps, items design, and outcome space, which are linked. The P-P-P Assessment's three holistic scoring guides, the backbone of its outcome space, systematically balance incorporation of planning, enactment, and reflection.

FARROP's rubrics, in contrast, are only focused on enactment. For FARROP, the planning and reflection aspects of teacher practice of formative assessment are part of the protocols that surround and support the observing and debriefing of instruction with teachers; planning and reflection are not explicitly part of the rubrics. In FARROP, planning and reflection are not embedded within the construct formative assessment. In the P-P-P Assessment, planning, enactment, and reflection are.

Outcome Space Design for the P-P-P Assessment

This section contextualizes, identifies and explains key design decisions in the creation of the outcome space for the P-P-P Assessment and describes the resulting iteration. Three holistic scoring guides, one each for the constructs of posing, pausing, and probing, were created for the P-P-P Assessment. In drafting and revising the outcome space, I aimed (a) to reflect qualities and procedures advocated by Wilson's "building

blocks" approach to constructing measures and (b) to ensure that it helped to generate meaningful, "next steps" feedback to teachers. Outcome space design decisions accounted for the expectation that future applications of the posing, pausing and probing scoring guides will include contexts outside of middle school mathematics instruction.

Design decisions. Designing an outcome space entails making choices specific to the construct being assessed and the contexts in which assessment of that construct will be used (Wilson, 2005). Design choices should work to strengthen inferential links between the outcome space and the construct maps and items design employed in the creation and administration of the assessment under development (Duckor, Draney, & Wilson, 2009; Wilson, 2005). This is so that the elements of the assessment can work in concert for an intended purpose as experts in the development of educational assessments affirm, as a necessity for good assessment designs (Pellegrino, Chudowsky, & Glaser, 2001).

Decisions following Wilson's approach reflect first principles of assessment. Highlighting the correspondence between Wilson's "building blocks" framework for constructing measures and first principles of the design of educational assessments warrants doing (again) here. Doing so underscores how design decisions made to follow Wilson's recommended approaches in creating an outcome space align with the recommendations of a large, recognized and respected body of assessment experts, the NRC's Committee on the Foundations of Assessment.

This committee's report (2001) represents the foundational elements of an assessment and their relation to one another through its "assessment triangle" and its three vertices of *cognition*, *observation*, and *interpretation.* All three elements should function in

synchrony to support (a) the purpose of the assessment and (b) arguments for the validity and reliability of the inferences that can be drawn from it when it is used as intended (Pellegrino, Chudowsky, & Glaser, 2001).

Each of the first three "building blocks" in Wilson's Constructing Measures framework (2005) —*construct maps*, *items design*, and *outcome space*—relates directly, in turn, to a vertex of the NRC's assessment triangle. Construct maps visually represent the theory of cognition—and hypotheses of progress—related to the latent construct being assessed. Items design functions to generate relevant observations. The outcome space design determines how the observations are to be interpreted in light of the theory of cognition and the purpose of the assessment. Most important to note in this chapter is that outcome space design choices should work in synchrony with choices regarding the construct maps and the items design in order to serve the assessment's intended purpose, while recognizing the inferential nature of the evidence-based arguments involved.

Examples of outcome space design choices. Outcome space design includes making decisions about what "counts" as a response, what can be scored, and what should be scored. Designing an outcome space requires deciding what will be ordered as more to less or higher to lower, and how the ordering will be defined. All these decisions should be made with the purpose, contexts, and targeted construct of the assessment in mind. Moreover, an assessment designer's understanding of the particulars and implications of such decisions is required in order to explain with authority where and when the outcome space design is likely to strengthen or weaken inferential links between specific aspects

of the "building blocks" in Wilson's framework when using the framework to inform the design of an assessment (Duckor, Draney, and Wilson, 2009).

What should be scored. The study's corpus of data—all the responses generated from the items design of the P-P-P Assessment—encompassed a sizable amount of content in several formats: survey responses (which included Likert-style and open-ended questions), transcripts of interview question responses, lesson plans, concept maps drawn by study subjects, video recordings of their enacted lessons, and lesson artifacts. All responses were potentially scorable and could figure to differing degrees in an outcome space.

Illustrating with an example: choosing not to score the Likert-style responses. The following example helps to illustrate how the intended purposes of an assessment, in relation to the target construct, can inform outcome space design choices. The outcome space design for the P-P-P Assessment could have included the scoring of the responses from the Likert-style items from the two surveys subjects took. This likely would have entailed employing the customary scoring scheme for such item responses: scoring them according to the number of response categories assessment takers were allowed.

For the P-P-P Assessment, this would have meant scoring *strongly agree*, *agree*, *disagree*, *strongly disagree* 0, 1, 2, and 3, respectively (or perhaps 1, 2, 3, and 4). If a central purpose of the P-P-P Assessment had been to discover teachers' attitudes and beliefs toward posing questions, providing wait/think time (pausing), and probing students to elaborate on initial responses, then I might have designed the outcome space to include this scoring scheme.

Aligning choice to aims of the assessment. However, the two central aims of designing the P-P-P Assessment were to (a) be able to locate a respondent and responses on a theorized continuum of practice in the domain of formative assessment as defined by the instructional "moves" of posing, pausing, and probing, and (b) to generate meaningful and useful "next steps" feedback for each respondent. The feedback aimed to be related by empirical evidence to the respondent's location on the theorized continuum of practice and the respondent's particular manifestation of planning for, enacting, and reflecting on posing, pausing, and probing. Therefore, while recognizing that teacher beliefs and attitudes intertwine with teachers' professional practices and that these are critical to take into account in efforts to influence teacher practices (Borko & Putnam, 1996; Guerra & Nelson, 2009), and while acknowledging that this can be a relevant factor when deciding what "kinds of" feedback might be more effective to give teachers in some contexts over others, I chose a different approach.

Critical decisions 1 and 2: To create three general, holistic scoring guides. Given the early stage of this work—this study is just the beginning of an expected longer trajectory of work in this vein—and given my goal to align the design of the outcome space with the purposes of the P-P-P Assessment identified ("a" and "b"), I needed to prioritize my efforts. Over other possible approaches, I prioritized considering how all the responses—all the study data—could inform the drafting and revision of three general, holistic, scoring guides. The prioritization of the drafting and revision of the three general, holistic scoring guides—one scoring guide each for posing, pausing, and probing—as the backbone of the P-P-P Assessment's outcome space is its most critical design feature. The three general, holistic scoring guides are meant to be applied "equally" usefully to evidence of teacher planning for, enactment of, and reflection on the moves. The decision to create only three general scoring guides—and not any item-specific scoring schemes, nor any item format-specific scoring schemes—is the second most critical design feature of the outcome space.

Implication: working toward application of each guide across performance task responses. For the P-P-P Assessment, this meant not designing a scoring scheme—such as by drafting a scoring guide or an analytic rubric—specifically to interpret responses to each of the three performance tasks of the P-P-P Assessment. The three tasks were (a) to plan an FA moves-infused lesson using the lesson planning template known as the "FA Moves Lesson Planning Tool," (b) to enact the lesson, and (c) to reflect on the enacted lesson through the lens of the FA moves and, in particular, on posing, pausing, and probing by means of a video-stimulated recall protocol.

This meant not creating a scoring scheme as part of the outcome space specifically for responses to the "FA Moves Lesson Planning Tool." Instead, I worked to ensure that each of the three general scoring guides included enough well-ordered categorization and description regarding lesson planning that the act of considering responses to the "FA Moves Lesson Planning Tool" in relation to any of the scoring guides could be a meaningful experience. Working this way supported my prioritization to consider how all the study data could inform the drafting and revision of three general, holistic scoring guides.

Implication: Structure of scoring guides. The commitment to create three general scoring guides that could be applied meaningfully to any subset of study data had implications for the structure of the scoring guides. All three share a common underlying structure related to the facets of FA practice—planning, enactment, and reflection—hypothesized in this study. Each scoring guide was structured to systematically balance descriptions related to planning, enactment and reflection of the move throughout its ordered categories. This was a direct outgrowth of this commitment.

Implication: Experience using scoring guides. At the same time, this choice suggests that if, for example, a rater were using a scoring guide to score responses to items that were planning-focused and the responses presented only planning-related evidence of the move; then most of the content of the scoring guide would not apply to that exercise, i.e., the content related to enactment and reflection. That choice, therefore, has implications for the use of the scoring guides. This could potentially influence tests of the inter-rater reliability of the scoring guides. This could happen in situations where the scoring guides

are applied to responses—whether by design or by happenstance due to the quality/characteristics of particular responses—that only feature—or predominantly feature—evidence of only one of the three facets of FA practice.

For example, responses to the "FA Moves Lesson Planning Tool," a lesson planning template, inherently would not feature observations of either enactment or reflection on a move. This could influence a rater's use of the scoring guide such that it might negatively influence inter-rater reliability. If one rater was especially challenged by the need to "skip" over the majority of the content of the scoring guide—as would be required in this situation—so much so that it negatively affected application of the scoring guide to the evidence; and another rater was not, the discrepancy in experiences using the same guide on the same evidence could decrease the rate of agreement between the two raters. The disagreement might be more a function of the how the content of the scoring guide is structured in the outcome space than it would be about the actual content itself.

SOLO taxonomy approach to drafting of scoring guides. As previously mentioned, the characteristics required for a sound and useful outcome space are that its categories are well-defined, finite and exhaustive, ordered, context-specific, and research-based (Masters & Wilson, 1997; Wilson, 2005, p. 64). As also previously mentioned, assessment designers using the building blocks approach frequently construct an outcome space using three general approaches: phenomenography, the Structure of Observed Learning Outcome (SOLO) taxonomy, and Guttman and Likert-item scales (Wilson, 2005).

Given the purposes and contexts for this assessment, I chose the SOLO taxonomy

approach for four reasons. First, the SOLO taxonomy is a general theoretical framework useful and appropriate for constructing an outcome space related to cognition (Wilson, 2005, p. 75) and therefore a match for an assessment focused on teachers' planning, decision making, and thinking/reflection related to posing, pausing and probing. Second, a "great strength" of the SOLO taxonomy is "its generality of application" (Dahlgren, 1984), also making it a fitting choice for an assessment that expects, longer range, to be applied to a wide variety of teaching contexts, wider than the specific context of middle school mathematics instruction common to all the subjects in this study. Third, the SOLO taxonomy has been used effectively in educational assessments for many years; it has a proven track record. Fourth, while assessment experts are alert to weaknesses of the SOLO taxonomy (Dahlgren, 1984, Duckor, 2006; Wilson, 2005), they also recognize its use as a sensible and strategic choice in many assessment contexts, especially as a way to get started (Wilson, 2005). For these reasons, I designed the outcome space for the P-P-P Assessment using the SOLO taxonomy approach.

A generalized SOLO taxonomy has five levels. From the lowest to highest, the ordered categories describing performances in response to a particular task are: pre-structural, unistructural, multistructural, relational, and extended abstract. Experienced users of the SOLO taxonomy have observed that employing this approach and its five levels can mean that the multistructural level "tends to be quite a bit larger than the other levels," (Wilson, 2005, p. 78). Table 14 defines each of the levels of a generalized SOLO taxonomy (Wilson, 2005, p. 75).

Table 14: *The SOLO Taxonomy*

Ordered category	Response
Extended abstract	is one that not only includes all relevant pieces of information, but extends the response to integrate relevant pieces of information not in the stimulus.
Relational	is one that integrates all relevant pieces of information from the stimulus.
Multistructural	is one that responds to several relevant pieces of information from the stimulus.
Unistructural	is one that responds to only one relevant piece of information from the stimulus
Pre-structural	is one that consists only of irrelevant information.

Note. (adapted from Wilson, 2005).

Early scoring guide (pausing) from initial outcome space design. Figure 12 presents an early version of the pausing scoring guide drafted before empirical evidence from the study had been analyzed and influenced revisions to the pausing construct map. Therefore, while the five categories in this version of the pausing scoring guide do relate back to the responses side of the generating construct map, as must be done when employing the "building blocks" framework to design and develop an assessment (Wilson, 2005, p. 69), the definitions of the categories and the level of interpretable detail associated with each category, do not reflect decisions to the design of the outcome space that considered complexity of the pausing construct only understood after examination of the empirical evidence.

5	***Contextualized and differentiated pausing (Think time+)*** ● Reflects purposeful attention to kid, curriculum, and item/task/prompt ● Exhibits a variety of pausing-related moves tailored to individual and group learning needs ● Supports more than students' "think time"
4	***Multi-faceted pausing (think time)*** ● Demonstrates different kinds of pausing for different classroom situations ● Values protecting "adequate" amounts of think time, including for self ● Serves to increase chances/opportunities for improved sampling of student thinking (elicits larger student response sample)
3	***Intentional pausing (wait time)*** ● Is backed up by reasons for pausing ● Lasts from a few to several seconds long ● Primary concern may appear to be getting everyone to quiet ● May display blanket approach
2	***Unsupported pausing*** ● Suggests that pauses that do occur happen "by accident" ● May be intentionally "ended" by teacher ● Undermined by teacher discomfort with silence
1	***No pausing***

Figure 12. Early version of scoring guide for pausing at beginning stage of the outcome space design for the P-P-P Assessment.

As well as presenting the content of the pausing scoring guide, Figure 12 shows the level of detail provided for each category in this version. In Figure 12, note that for three of the four categories that possess explicating detail on the pausing scoring guide—levels 2, 4, and 5—(level 1, as a category defining "no pausing," was determined to need no explicating detail)—three bullet points of text accompany the main descriptor of each category. The main descriptors of each category appear in bolded italics in Figure 12. Level 3, which represents mid-range of the continuum of practice described by the

outcome space, possesses four bullet points.

As empirical evidence influenced my understanding of the pausing construct, and I revised the pausing construct map to reflect this understanding, I necessarily re-considered my initial design for the outcome space for the P-P-P Assessment. This included revising the early version of the pausing scoring guide shown in Figure 12 and the early versions of the posing and probing scoring guides too. Overall, the revisions increased the level of detail that the scoring guides provided for each category they defined. Representative of this increase in detail, the number of bullet points of text describing salient, distinguishing characteristics of respondent performance at each level of the pausing scoring guide went from 0 to 3 at level 1, from 3 to 7 at level 2, from 4 to 7 at level 3, from 3 to 5 at level 4 and from 3 to 4 at level 5. These revisions sought to strengthen the inferential links between the set of three construct maps and their associated scoring guides in the design of the P-P-P Assessment.

Aiming for "sufficiently interpretable detail". An outcome space should provide "sufficiently interpretable detail" (Wilson, 2005, p. 65) such that different teams of assessors considering a common set of responses in light of the categories defined by the outcome space will agree on (a) the number of categories used to define the outcome space, as well as (b) the essential elements of those categories as communicated by the descriptors and explicating details. In the context of further development of the P-P-P Assessment, that would mean different evaluators applying the posing, pausing, and probing scoring guides to the lesson plans, video clips of lesson enactment, and oral responses to the video-stimulated recall protocol of the teachers who took P-P-P

Assessment. This process has not yet been conducted, however, given the aims and scope of this study.

Figure 13 presents an exemplar scoring guide from the P-P-P Assessment's outcome space after revisions. Empirical evidence from the study and feedback from a panel of expert evaluators of mathematics teachers inspired the revisions to the design of the outcome space for the P-P-P Assessment. Figure 13 represents the final version of the pausing scoring guide for this study (see appendix B for all three final versions).

5	**Contextualized use of "think time" based on curricular challenge and/or student learning style/need** Teacher/instruction/pausing ● is tailored to individual and group needs (e.g., ELs, students with 504 plans) ● reflects purposeful attention to student, curriculum, and task/prompt in relation to learning target ● plans for and can explain why several different kinds of pausing moves are used with which students, why, and when in the lesson and learning cycle ● reflects on how pausing practices could better serve individual/group needs
4	**Strategic use of "think time" improves student access to curriculum and teacher decision making** ● includes a mix of "quiet", "noisy", "active", "still", "individual", "group" "directed" and "undirected" pausing routines selected to fit learners' needs regarding advancing toward lesson target ● makes own needs for pausing for "think time" a priority ● encourages students' roles/responsibilities regarding pausing ● can explain several benefits of structured pausing and how pausing can influence decision making ● unpacks practices related to pausing from more than one orientation (e.g., learner-focused orientation, assessment-focused, equity-focused)
3	**Intentional use of "wait time" includes routines for non-silent pausing** ● features "pair-shares" and "table talk" as "go to" pausing moves ● demonstrates verbal and nonverbal support of pausing ● whole class silences last from a few to several seconds long ● may include pausing routines inserted on the fly when "not enough hands up" or "too many blank looks" ● pausing moves may appear "one size fits all" ● plans for pausing moves to increase student participation and elicit "better" student responses, though may not be well-articulated on planning

	documents ● offers suggestions/reasons for improving pausing
2	**Concerned with getting self and students quiet for lengths of time** ● Unplanned, spontaneous pauses "prematurely" ended by teacher ● features no public expressions during class of valuing pausing or wait/think time ● may include intermittent pauses that happen incidentally/accidentally ● plans lessons that do not include explicit pausing procedures ● is undermined by teacher discomfort with silence, classroom management skills, or (lack of) confidence in management skills ● planning documents do not anticipate places in the lessons where pauses are needed to support student learning, promote more equitable participation, and increase quality of responses ● reflection may explore reasons for "rushing," may include suggestions for pausing, often tactics (e.g., "I will count.")
1	**No pausing moves are planned or observed during enactment** Teacher ● identifies and reflects on missed opportunities for pausing ● may acknowledge importance of pausing ● may offer ideas/suggestions for how to support/improve pausing

Figure 13. Scoring guide for pausing, end of study.

Note that though the number of categories has not changed in the two versions, the level of detail provided has increased in every category defined. The greatest increase in detail between the two versions of the pausing scoring guide occurred in the middle three categories defined—levels 2, 3, and 5. This befits the study's focus on how the latent characteristics (or constructs) of interest—teacher posing, pausing, and probing—can be manifested in the range of practice of formative assessment between novice and expert. This increased level of detail at these three middle levels in particular also potentially serves the central purpose of the P-P-P Assessment better than the previous level of detail provided. That purpose is to allow for and support the generation of meaningful, "next steps" feedback to teachers completing the Assessment.

Strengths and limitations. Arguably, the planning, enactment, and reflection practices of most classroom teachers in the field would locate them most readily within the middle three levels of the P-P-P construct maps. This suggests a certain kind of potential regarding the usefulness of the P-P-P Assessment's outcome space.

The scoring guides of the P-P-P Assessment have been related back to the construct maps in such a way that they support the generation of "next steps" feedback to teachers. Sample feedback to teachers who took the P-P-P Assessment is explored next in chapter 5, "Profiles of Practice and Feedback."

The P-P-P Assessment stands out as an instrument in K-12 that can help to generate meaningful "next steps" feedback to teachers on the planning, enactment, and reflection of essential formative assessment practices. This is due to the key step taken during the development of the P-P-P Assessment's outcome space of relating the ordered categories back to their generating construct maps, and doing so in light of empirical evidence. Though every effort has been made to provide sufficiently interpretable detail in the P-P-P Assessment's scoring guides, until different evaluators, or different teams of evaluators score the same body of work with the scoring guides—or until assessment moderation (Wilson & Sloane, 2000) is conducted—it will be impossible to determine the success of the definition of the outcome space.

Chapter 5: Profiles of Practice And Feedback

This chapter presents profiles of the posing, pausing, and probing practices and planning, enacting, and reflecting skills of the six teachers in the study. These evidence-based descriptions of the teachers' practices in these three dimensions and three facets of formative assessment were compiled from the teachers' responses to the P-P-P Assessment. For this study I used the evidence to locate respondents and their responses to the items/tasks of the P-P-P Assessment on the finalized versions of each of the three construct maps drafted and revised, based on empirical evidence. Locating each respondent and their responses to items/tasks of the P-P-P Assessment functions as an evaluation of their practices of posing, pausing, and probing moves along the generalized continua articulated in this study. In doing so, I addressed aspects of each of the study's three research questions that were presented in chapter one. First I describe the structure of the profiles, the organizing principle behind their content and the order of their presentation. Next I present the six profiles themselves.

Structure of the Profiles

Each profile begins with a description of relevant background information and context concerning the respondent and responses evaluated. This includes information about each respondent: teaching experience, credential or credentials held (single and multiple subject), content knowledge, engagement with general and mathematics-focused professional development, and length of experience with the curriculum. The beginning of each profile also communicates contextual information on the class and the lesson for which, in which, and about which (planning, enactment, and reflection) the evidence-

gathering occurred. This contextual information includes: length of class period (block or regular), numbers and selected demographics of students in the class, unit in which the lesson occurred, lesson objective or objectives, and mathematics and English Language Development standards planned to be addressed in the lesson. In every case the teacher chose which class and lesson she or he used to respond to performance tasks one, two, and three of the P-P-P Assessment, i.e., (1) plan, (2) enact, and (3) reflect on a FA moves-infused lesson.

Location by dimension and facet. The rest of each profile adheres to the structure I describe next. Each dimension of FA practice—posing, pausing, and probing—was taken in turn. For each dimension, first an overall assessment was given. The overall assessment answered: (a) based on evidence, where on the five levels on each construct map was this respondent and their responses to the P-P-P Assessment located? and (b) what does this mean?

Next, each profile communicates where the responses of each respondent locate the respondent's planning for, enactment of, and reflection on that dimension. This portion, which comprises the majority of the content of each profile, identifies and presents (a) the specific evidence the resulting location on the relevant construct map was based upon and (b) how that evidence relates to that construct map. As the items design of the P-P-P Assessment intended, and as respondents' responses allowed (e.g., two respondents did not take the pre-lesson enactment survey; therefore these two teachers' self-report evidence from their responses to survey items were not available to me), I triangulated data from multiple sources to determine the location on the construct map, i.e., the level

of each teacher's skill in planning for, enactment of, and reflection on posing, pausing, and probing.

Individual formative feedback. The locating-based-on-evidence portion of each profile is followed by a succinct articulation of formative feedback for the individual. The feedback is tied to the responses generated from each teacher's engagement with the P-P-P Assessment and the resulting evaluations of their posing, pausing, and probing. The feedback identifies areas for growth related to the three dimensions and three facets of FA practice that were the targets of investigation of this study. The individualized formative feedback is intended to fall within each respondent's zone of proximal development in the practice of formative assessment as conceptualized in this study.

Order of Presentation

I present the profiles by school in alphabetical order: first Chavez, then Kimm, then Sierra Middle School. There were two teachers at each school in my study. For each school I present the teacher with fewer years of teaching experience first, the teacher with more years of experience second. This order of presentation reflects a core purpose behind creating the P-P-P Assessment: to be able to provide teachers at any school (who have various years of teaching experience and who possess varying levels of proficiency in the practice of formative assessment) with meaningful formative feedback. The profiles are *not presented* in a ranked order of levels of practice as determined by responses to the P-P-P Assessment.

Leila

Background and context. A former engineer with a bachelor's in electrical engineering, teaching is a second career for Leila. During the study, Leila was in her third year teaching middle school and her second year teaching seventh grade math and science at Chavez. English is her second language.

Leila holds a multiple subject credential (MSC). She completed the MSC program at a local state university, including the Performance Assessment for California Teachers (PACT). Her work for the PACT was Leila's only other experience of reflecting on video clips of her instruction.

Leila had never been a member of a math-oriented professional organization, such as the National Council of the Teaching of Mathematics (NCTM). Leila had attended three meetings sponsored by a nonprofit organization dedicated to catalyzing changes in local math instruction so that students' math achievement might rise.

It was Leila's second year using the College Preparatory Mathematics curriculum (CPM). She had attended four district-sponsored meetings about instructional strategies promoted in the CPM curriculum, such as "giving guiding questions" as students worked to come up with solutions "whether they're right or wrong." CPM curriculum supports cooperative learning. Leila's students regularly served in CPM-supported roles such as resources manager, reporter, task manager, or facilitator. Students did not serve in these roles during the lesson video recorded for this study.

Leila chose to plan, enact, and reflect on a lesson for an integrated math and science class that met daily for 90-110 minutes. Leila reported "I'm trying to integrate science,

but my priority is to make sure that my students meet the Common Core math standards." No science content appeared in the video recorded lesson.

Students in Leila's class reflected the demographics of the population at Chavez: 85% Hispanic or Latino and 11% Asian and Filipino. Ninety-five percent were officially categorized "socioeconomically disadvantaged." Forty-four percent were designated English learners. Of 24 students in Leila's focal class, 10 were designated English learners. Three had 504 plans.

The lesson occurred during a unit on proportional relationship and expected students to create tables and graph proportional relationships. Students were expected to know from previous lessons that two quantities are related proportionally if the ratios of these quantities are equivalent at any given time. On the lesson planning template, for "Background" Leila wrote "Students learn that a proportional relationship is multiplicative and is not additive." The student learning objective (SLO) was "SWBAT verbally explain how the relationship of two quantities is proportional, and show the proportional relationship of two quantities by plotting a linear graph that goes through the origin using data from a given table."

According to Leila's lesson planning template response, her lesson was targeting two seventh grade Common Core State Standards in mathematics (CCSSM) on ratio and proportion (RP):

> 7. RP. 2a) Decide whether two quantities are in a proportional relationship, e.g., by testing for equivalent ratios in a table or graphing on a coordinate plane and observing whether the graph is a straight line through the origin.
>
> 7. RP. 2d) Explain what a point (x ,y) on the graph of a proportional relationship

> means in terms of the situation, with special attention to the points (0, 0) and (1, *r*) where *r* is the unit rate.

The lesson aimed to address ELD standards SL7. 4-6 and L7. 1, 3, which Leila conveyed as "Expressing information and ideas in formal oral presentations on academic topics."

Posing practice. Overall, evidence of Leila's planning for, enacting, and reflecting on posing generated by her engagement with the P-P-P Assessment located her practice at the multistructural (3) level, constrained posing. Respondents exhibiting multistructural (3) level, constrained posing, seem to be trying to get students to say what they, as the teacher, are thinking rather than eliciting from students a range of responses, including unknown responses and responses surprising to the teacher. At the multistructural (3) level, questions planned connect to the learning target. Most, or even all questions, posed during enactment are fact recall—what, when, where—and lower-level questions according to taxonomies such as Webb's, Bloom's, and Costa's.

Planning. Leila's response to the lesson planning template demonstrated that she planned questions to serve as checks for understanding of the lesson's objective, which aligns with expectations of planning for posing at the multistructural (3) level. Two such questions on her lesson plan were: (a) "What does it mean for a relationship to be proportional?" and (b) "Where on the graph can we show the cost for buying nothing?" Both were tied to the following SLO of two SLOs Leila listed on her lesson plan: "plotting a linear graph that goes through the origin using data from a given table." This evidence generated from Leila's response to the lesson planning template indicated her planning for posing as illustrative of multistructural (3) level posing practice.

Leila's response to the lesson planning template also revealed that she planned a routine around one of the key questions for the lesson. She planned for two students to engage in Reciprocal Teaching pairs as they considered "What does it mean for a relationship to be proportional?" This is a strength of Leila's posing that matches the part of the description for relational (4) posing on the posing construct map that reads, "Respondents plan supports/scaffolds for questions." But it is not enough to locate Leila's planning for posing at the relational level.

There is also evidence from Leila's response to the lesson planning template demonstrating that Leila was anticipating where she thought students would become confused and or stuck and was planning questions toward those areas. The questions were (a) "What is the cost to you if you do not buy anything in a store?" and (b) "Where on the graph can we show the cost for buying nothing?" They reflect Leila's anticipating (a) that students would not see that they *need a data point at the origin*, and (b) might not readily understand what "at the origin" means in the real-world scenarios they were working with and attempting to graph. Leila anticipated that students would have difficulty realizing that "buying nothing" or "buying no pounds of cheese" in a scenario about cheese that costs $2.50 a pound means graphing a point at (0,0), the origin. This sort of anticipation of student thinking by a teacher is characteristic of extended abstract (5), or integrative posing. By itself, however, this evidence from Leila's response to the lesson planning template is not enough to change her location regarding planning for posing from the multistructural level.

Enactment. Leila's posing enactment was located at the multistructural (3) level, constrained posing. Though posed questions observed during lesson enactment and captured on video were tied to the objective of the lesson, analysis of video and transcript evidence revealed the questions did not elicit a wide range of responses. For example, toward the end of the lesson during whole class configuration, Leila asked, "Which part specifically on the graph must it [the graphed line] go through in order to show those two quantities to be proportional, Esteban?" Transcription of the video reveals Leila posed six very similar versions of this question in rapid succession: her pose to Esteban in particular and five poses to the whole class. All together this quick series of questions elicited three qualitatively different responses. They were: (a) "all parts of the graph", (b) Esteban's response, which was "The pound", and (c) "zero" (the answer Leila was looking for) —not a wide range.

Analysis of video evidence also reveals that throughout the lesson, during both small group and whole class configurations, and indicative of multistructural (3) level posing, Leila asked a high percentage of lower-level, closed-ended questions. For example, eight of nine questions Leila posed while interacting with one small group were close-ended, lower-level questions according to Bloom's taxonomy. This was representative of her posing during that lesson and supports locating the enactment facet of her posing practice at the multistructural (3) level.

Reflection. Leila's reflection on her posing explored specific ways of improving her posing, indicating multistructural (3) level reflecting on posing. For example, during VSR, Leila saw a mismatch between how she worded one of her poses to check for

understanding and the graphic representations the students were using. Leila hypothesized that this mismatch affected students' responses to her posing. Leila's question was intended to be a straightforward check for understanding about the necessity for linear graphs of proportional relationships to pass through the origin (0, 0). Leila asked the class, "Where does the line go through?" When no student responded with the answer she expected—the origin, Leila asked the question thrice in a row.

Reflecting, Leila noted all the graphs the students had used during the lesson and the graph she was pointing to as she posed this question showed not all four quadrants of a coordinate plane with a line intersecting or "going through" the origin, but only quadrant I. The graph showed a "partial" line that seemingly "began" or "started" at the origin and continued into quadrant I. None of the graphs students had interacted with during the lesson had shown more than quadrant I. None—including the graph students were looking at as Leila posed the question—showed a "complete" line "going through" the origin to extend into quadrant III.

Leila's reflection demonstrated she was able to see the lesson from her students' point of view. During VSR, Leila said, "The language for the definition of proportionality always says the line *passes through* the origin. But it doesn't on their graph. Nothing goes 'through' anything. To them, it's like, 'I didn't see anything that passes through the origin.'" The transcript of Leila's response to the VSR protocol reveals Leila went on to suggest two ways to improve her posing related to that situation, indicating that evidence supports locating Leila's reflecting on posing at the multistructural (3) level.

Pausing practice. Overall, evidence of Leila's planning for, enacting, and reflecting on pausing located her practice at the unistructural (2) level, unintentional, unsupported pausing. Respondents at the unistructural level want to avoid stretches of silence and may feel fear, distrust, discomfort, or pain about silence in class, especially in response to questions they have posed or directions they have given. Responses to items/tasks indicate that pauses that do begin spontaneously are ended prematurely by teacher action that interrupts the pausing. At this level, teachers tend to miss cues that pausing is needed. With unintentional, unsupported pausing, student discipline or time on task can confound with the construct. Classroom action related to pausing is not well organized or modulated at this level of pausing practice.

Planning. Leila's response to the lesson planning template generated evidence that supports locating Leila's planning for posing at the unistructural (2) level. With unistructural pausing practice teachers plan lessons with few or no explicit pausing moves or routines. Leila planned for students to do a "Think-Pair-Share" in the first component of her lesson. She identified this component the "Engagement" component on her lesson plan. This was the only explicit routine to support pausing evidenced in Leila's lesson plan. In responses to the pre-lesson enactment survey items, Leila self-reported she "often" planned pausing routines (response to item 40), but that when students need 'wait-time' she "rarely" creates specific routines for them (response to item 41). Consideration of this self-report evidence does not change locating Leila's planning for pausing at the unistructural level.

Enactment. Though Leila's response to the lesson planning template showed that she had planned to enact a Think-Pair-Share with her students, live observation by the researcher, video recording of the lesson, and transcripts of the video recording revealed that no routines to support pausing (whether quiet pauses or non-silent, "talking" pauses) were enacted before or after Leila posed questions. Analysis of the evidence from these same data sources additionally revealed no references to wait/think time during lesson enactment. After Leila posed a question, no pauses beyond two seconds were observed in the video recording of the lesson. Transcriptions of video recordings also included time markers. Students either called out responses immediately or Leila re-posed the question. There was one exception, which was illustrative of an unplanned spontaneous pause being "prematurely" ended by the teacher, which is indicative of unistructural (2) level practice of pausing. In this one instance of a pause lasting longer than two seconds, the silence after one of Leila's questions lasted three seconds, at which point Leila ended the pause by posing a different question. Analysis of evidence of Leila's lesson enactment locates the enactment facet of her pausing practice at the unistructural (2) level, unintentional, unsupported pausing.

Reflection. Leila's reflection on her pausing, generated in response to the VSR protocol, did not explore reasons for rushing, or reflect on specific missed opportunities for pausing/pausing moves. Nor did it include suggestions to improve pausing or offer next steps to try to improve pausing as unistructural (2) reflection on pausing characteristically includes, according to the current versions of the pausing construct map and scoring guides. During VSR, however, Leila did talk about why pausing is important

for students, an essential characteristic of the reflection facet of pausing practice at the unistructural (2) level.

Analysis of the transcript of Leila's VSR protocol responses show she reflected on why a certain type of non-silent pausing is important for certain students. The certain students were her "focal students." The type of non-silent pausing shall be referred to as "re-visit pausing" since, during VSR, Leila identified the time between her "re-visits" to an individual student or to a group of students as pausing. Leila explained she enacted this kind of pausing for her "'focal students,' who need extra time to think it through, they are struggling." Leila explained while reflecting on her video clips of her lesson enactment that the purpose of one of her "re-visiting" pausing moves was to "give the students time to work together and talk about what they think of the quantity and what they see the graph represent."

Video evidence captured Leila enacting this "re-visiting" pausing with Ricardo, who happened not to be one of Leila's focal students. Ricardo was "a high-performing student" according to Leila. During VSR, Leila reflected:

> I am waiting for him [Ricardo] to figure it out. I am waiting. I have given him a lot of guidance and then...I stop talking. I have him figure it out. I went back. I told him [Ricardo], "I'll give you some time to process it. When you're done, you come back [to me]. When I pause, if the student is willing to think and struggle, then they will do it.

In response to the VSR protocol, Leila spoke about conducting her "re-visiting" pauses differently with her focal students than with "students like Ricardo." Leila said, "I will check [proactively] on the work of my focal students." Leila's reflection on pausing, since it explored reasons for such pausing, (e.g., "[struggling] students need extra time to

think it through" and "I am pausing for those focal students and I will check on their work") located her reflection on pausing at the unistructural (2) level.

Probing practice. Overall, evidence of Leila's planning for, enacting, and reflecting on probing located her practice at the multistructural (3) level, targeted probing. Respondents exhibiting targeted probing distinguish between probing to assist teacher decision making and probing to benefit student(s) being probes or any witness learner(s) to the teacher-student(s) probing interaction. At this level, responses to items/tasks indicate that probing is potentially valuable to teachers or student decision making. Probes target uncovering misconceptions. The teacher leverages probes and what probes make visible at this level of probing practice.

Planning. Leila's planning for probing, as evidenced in her response to the lesson planning template, demonstrated planning specific probes for different points in the lesson, indicating this facet of her probing practice is located at the multistructural (3) level— targeted probing—on the continuum of probing practice articulated in this study. Leila's response revealed that she had planned specific probes for the "Engagement" and "Instruction" components of her lesson.

Enactment. Enactment evidence—video recording of the lesson, researcher observation of the lesson, and transcripts of the recorded video—indicated multistructural (3) probing. Evidence showed that during the lesson Leila leveraged the information that the probes made visible. For example, after asking a series of probing questions to a small group of students, Leila was video recorded saying to them, "I know what the confusion is." Leila proceeded to help the students make sense of two graphs they were

comparing. In one graph the intervals depicted were by units of 100, but in the other graph, they were units of 1. The graphs were the same size in their textbooks. The distance between hash marks that indicated units on the axes of both graphs was the same on both graphs too. After probing the students and thinking about how students had responded to her probing, Leila spoke with the students about "interval" and its meaning in relation to the two graphs. She leveraged what probing had made visible to her.

Analysis of video evidence of Leila's enactment also demonstrated Leila attempting to advance the class's understanding of the learning target by using what was elicited via probing, which is indicative of multistructural (3) practice of probing. For example, toward the end of the lesson, Leila used data she had elicited from a student named Anastasia to attempt to advance the class's understanding of (a) proportional relationships, (b) that the quantities involved in proportional relationships are multiplicative in nature, and (c) the concept and term, *constant*. Leila's probing of Anastasia's work process elicited that Anastasia had "multiplied everything by 3." Leila then used this elicited response to attempt to advance the class's understanding by asking the entire class, "Anastasia said she multiplied everything by 3; what would that 3 represent?"

Enactment evidence also revealed that a focus on probing depth sacrificed achieving a wide range of information via probing, and constrained the evidence available to inform Leila's decision making as it related to a snapshot of the whole class's understanding pertaining to the learning target. This is indicative of multistructural (3) probing. Analysis of video evidence reveals that Leila probed one student's thinking, Ricardo's, more times

than any other student's thinking. By lesson's end, evidence suggested that Leila knew much about Ricardo's thinking and his solution processes to the tasks given. This included knowing about Ricardo's having "experienced" a proportional relationship as multiplicative and not additive, since Leila was standing next to Ricardo and probing him as he had this "experience." But video evidence and direct lesson observation also showed that Leila had not questioned many other students besides Ricardo about their recognition of this property of proportional relationships, which was one of the learning targets Leila had established in her lesson plan: "Students learn that proportional relationship is multiplicative and is not additive."

Reflection. Analysis of Leila's response to the VSR protocol revealed that Leila's reflection on probing demonstrated that she was focused on formulating improvements to her probing that could benefit certain learners. This is indicative of multistructural (3) reflection on probing. In Leila's case, the "certain learners" were students "who think that adding is the right thing to do".

During VSR, Leila pointed out that during her probing, "I could have given an example that would have shown that addition does not show proportionality, but I just couldn't think of anything [in the moment, during the lesson]." Leila could not come up with an example during VSR ("I have to think about what kind of example," she said), but she did explain how she could use that example once she determined the right one to use for her purpose, which was to help the "handful of students still adding things when they're trying to find that relationship". Further, in her suggestions for improving her probing, Leila provided a degree of specificity characteristic of multistructural (3)

reflection on probing:

> During small group instruction I can pull up that example to show the difference between addition and multiplication, or which one is it, and then maybe we can actually graph something, but it should not be a straight line by adding things.

Being able to come up with improvements to her probing that could benefit certain learners during VSR locates Leila at the multistructural (3) level for this facet of probing practice.

Areas for growth. Overall, Leila's formative assessment practice would benefit from a focus on planning for and enacting pausing moves and routines to support student think time. Enacting supportive pausing routines should serve to increase the number of student responses (e.g., hands raised) elicited by Leila's posed questions during whole class instruction. This will support better (more representative) data collection when Leila poses questions as checks for understanding. Longer pauses (think time of more than 2 seconds) after whole class poses may give Leila time to be more intentional about which students she calls on. Leila's current pausing practice suggests that the voices of slower-to-respond students are not getting consistent opportunities to be heard in the question-and-answer exchanges that occur during whole class configuration.

Posing. Taking steps to elicit a wider range of responses (e.g., through priming and pausing moves tied to poses) from questions posed is an area for growth. A fruitful focal point related to posing would be to support more students to ask questions at several points in the lesson. Current practice suggests routines intended to foster students asking questions are not enacted regularly.

Pausing. An area for growth in Leila's pausing involves noting that during enactment, when pauses spontaneously occur after a posed question during whole class configuration, that Leila is the one to end the pause, and quickly (typically before three seconds have elapsed). Evidence suggests several benefits to pausing longer and that most students need more time than that to think before they respond.

Probing. Since no encouragement of student-to-student probing was evidenced, taking action to support student-to-student probing is recommended. Focusing attention on how much probing occurs during different components of the lesson is recommended. This sort of analysis can inspire the creation of methods for increasing the amount of probing that occurs during class time overall (e.g. by scaffolding and orchestrating student-to-student probing) and for coming up with new ways to probe more strategically.

Planning. Planning a variety of pausing moves for different points in the lesson is recommended as a next step to improve FA practice. Anticipating student questions during lesson planning is also recommended as a strategy to enhance the planning of probes.

Enacting. The area for growth necessitating the highest prioritization is enactment of pausing. Since stronger practice of posing, pausing, and probing was observed during small group and one-on-one configurations compared to whole class configuration, an area for growth is FA practice during whole class instruction. What might need to be in place for practices that are occurring during small group configuration to occur during whole class configuration?

Reflecting. An area for growth concerning reflection relates to the variety of explanations (grounded in evidence) offered to explain student actions and possibilities for next steps. Increasing the breadth of each would improve this facet of FA practice. Both explanations and next steps should (continue to) be grounded in evidence. Being encouraged to provide sound and defensible rationales for next steps as they are suggested should serve to improve the quality of reflecting.

Lavinia

Background and context. At the time of the study, Lavinia was in her twentieth year of teaching and her second year teaching eighth grade mathematics. Except for a year teaching at a charter school, all of Lavinia's teaching had been in her current school district teaching grades three, four, and five in self-contained classrooms and teaching combined history and English language arts in grade six.

Lavinia held a multiple subject credential and had been a dance major as an undergraduate. She had volunteered to switch from teaching sixth grade history and English language arts to teaching eighth grade math and science upon learning from her current principal of her school's need for an eighth grade math teacher. The district was transitioning to the CPM curriculum, so Lavinia said that it was a fortuitous time make the switch. She had been attending all of the district-sponsored CPM trainings offered to her.

Lavinia credited her brother, who had been a math major ("he's the mathematician in the family"), with helping her deepen her mathematics content knowledge. He had taught mathematics for a time. Lavinia reported that he had been a tremendous resource to her

during the previous year when she had been "only a step ahead of [the kids]".

Lavinia had 30 students in her block period integrated math/science class which met for approximately 105 minutes every morning. The demographics of the class reflected the demographics of the population of students attending Chavez: approximately 86% Hispanic or Latino, 10% Asian, 1% Filipino, 1% white, and 1% Black or African American. Forty-four percent of students school-wide were officially designated English learners (ELs). Fourteen students in Lavinia's focal class for the study were designated ELs. Three had 504 plans.

The lesson Lavinia planned, enacted, and reflected on for the study occurred during the unit "Finding and Understanding Patterns, Sequences, and How They Grow." The lesson was titled, "Finding Rules for the Patterns We See" and was intended to last two class periods. The lesson summary stated

> students will be creating team webs to come up with ways they think *discrete* and *continuous* graphs are useful. Then they will revisit looking at patterns and figure out how it is growing and changing. Whole class discussion to review vocabulary words and examples in order to get background knowledge.

The SLO was "SWBAT find meaningful ways that we use *continuous* and *discrete* graphing. They will be able to figure out how a pattern is growing and changing in multiple ways so they can ascertain what different figures are 1-100 based on the formula $mx + b$."

According to Lavinia's response to the lesson planning template, the lesson was targeting the Common Core State Standard in mathematics that Lavinia conveyed as "Students will be able to grasp the concept of a function as a rule that assigns to each

input exactly 1 output. They understand that functions describe patterns and situations where one quantity determines another." Lavinia's lesson planning response communicated the lesson was addressing this ELD standard: "Students will be able to explain and describe the patterns they see as they analyze mathematical shapes and how they change."

Posing practice. Overall, evidence of Lavinia's planning for, enacting, and reflecting on posing located her practice on the continuum above the multistructural level (3), constrained posing, but not solidly within or matching the description of relational level (4), flexible posing. Key characteristics of both levels of posing are synthesized here.

Lavinia's overall level of posing practice was above the multistructural level of posing. At the multistructural (3) level, teachers plan questions connected to the learning target and during these lessons, a high percentage of questions posed are lower-level questions according taxonomical questioning schemes such as Bloom's, or Webb's Depth of Knowledge (DOK). At this level of posing, the purposes for questioning often seem to be to get students to say what the teacher is thinking, rather than elicit a range of responses, including responses the teacher has not anticipated or cannot reasonably anticipate (e.g. unorthodox responses).

Though above the multistructural level of practice, Lavinia's posing could not be solidly located at the relational (4) level however. At the relational (4) level of practice in posing, respondents demonstrate skillful, strategic flexibility in their questioning. The purposes of their questioning include eliciting a wide range of responses. Particularly the purposes for their questioning include eliciting misconceptions about the learning target

and student responses they cannot anticipate. They ask *how* and *why* questions and questions designed to foreground students' metacognition. At the relational level of practice, their posing reflects questions *from a strategic mix* of levels within a taxonomy such as Bloom's, Costa's or Webb's.

Planning. Lavinia's response to the lesson planning template illustrated that Lavinia planned questions she considered checks for understanding of the lesson's objective, which is indicative of multistructural (3) planning for posing. In her lesson plan Lavinia wrote

> "CFU: I will go to each team to see what they observed and figured out. I will help them as needed by asking questions to help them come to their own conclusions: "Hmm. I'm not sure I understand this pattern part you made. Can you help me? Do you mean…?"

And though Lavinia's response to the lesson planning template demonstrated an awareness of the need to match questions to specific purposes as posing practice at the relational level (4) does, analysis of Lavinia's entire lesson planning response does *not* show evidence of other aspects of planning questions necessary to be located at level 4: (a) planning questions designed to elicit misconceptions and unorthodox responses, (b) planning carefully sequenced repetition of key questions, and (c) planning support/scaffolds for student questions.

At the same time, Lavinia planned the following questions to use with students as they worked in teams. They are evidence of Lavinia's planning questions beyond those questions she considers checks for understanding of the lesson's objective (level 3 planning for posing):

I want to challenge each team to finding at least two different patterns.

- How do you see this pattern growing?
- How is it changing?
- What stands out to you?
- Can you figure out what figure 5 and 6 look like? How do you know? Can you explain it to me?
- Does anyone see a different pattern?

Lavinia's planning for posing locates this facet of her practice above the description for the multistructural (3) level of posing, but not matched to the description for the relational (4) level

Enactment. Enactment evidence indicated that Lavinia's posing practice was beyond much of the descriptions of practice of multistructural level (3), constrained posing, but not well-matched to characteristics of relational level (4), flexible posing. Analysis of video revealed that half of Lavinia's poses during whole class configuration were open-ended questions. This would locate her posing practice above the multistructural level (3) in which a high percentage of teacher questions posed are lower-level and closed-ended. But for a teacher's posing practice to be located at the relational level (4) on the posing construct map, lesson enactment needs to provide evidence of posing's relationship to contingency: "activities and pacing clearly reflect teacher decisions that are contingent upon student responses to questions posed about the learning target." This was not found in analysis of directly observed or video recorded lesson enactment evidence.

Though observation of Lavinia's teaching did show the posing of *some* questions that served to highlight connecting students' prior knowledge and experiences with present efforts to engage with and "reach" the learning target, to support locating a teacher's practice at level 4, relational posing, the evidence relevant to the enactment facet of

posing practice needs to demonstrate that *many* questions posed by the teacher during the lesson do this. *Some* questions are not enough. Analysis of video evidence revealed Lavinia asking a student, for example, "Can you think of another reason why music might go under continuous graphing?" However, complete analysis of the transcription of the entire lesson shows there were not many such questions, as the description of posing practice at the relational (4) requires.

Reflection. Lavinia's reflection on her posing did not map well to the reflection-related descriptions on the posing construct map. Before presenting the evidence relevant to Lavinia's reflecting on posing, I present all the descriptions from the most current version of the construct map that pertain to reflection on posing. Since there are five levels of practice hypothesized on the construct map, there are five descriptions of reflecting on posing. They are:

> Respondents who:
> (1) are able to "reflect" through descriptions of their instruction that do not push to analysis (pre-structural or *pre-posing* reflection on posing).
> (2) are able to reflect on benefits that might accrue from using a questioning scheme (unistructural or *posing to manage* reflection on posing).
> (3) are able to reflect on several aims of improving posing. Reflection includes specific suggestions for alternate poses to try (multistructural or *constrained* reflection on posing).
> (4) are able to reflect on perceived effects of changing questions and/or questioning strategies. They are able to suggest several "next steps" likely to support improved posing. They do so from many perspectives and with specificity (relational or *flexible* reflection on posing).
> (5) are able to reflect on how questions posed functioned to elicit evidence of student understanding in relation to lesson objectives/target(s) of instruction (extended abstract or *integrative* reflection on posing).

Instead, during VSR Lavinia spoke to craft aspects of how she posed: how she stood,

crouched, or knelt in relation to a student as she asked a question; the volume of her voice; the rate of her speech; her eye contact; and her use of humor. For example, as Lavinia unpacked an episode, a video clip, of her interactions with a small group, she reflected on her intentional control of her body language and voice while interacting with a particular student she knew well, Gerald.

Gerald, according to Lavinia, "struggles". "When he gives a presentation, he trembles," Lavinia reported during VSR. "It's a really painful thing and I have made him do it anyway. I praise him. I try to be very careful when I praise so that Gerald doesn't get embarrassed."

Lavinia's response to the VSR protocol generated her explanation that she "drew close [to Gerald physically], more personalizing it, talking a little softer to match where he was at". Lavinia reflected, "You kind of have to make yourself smaller and slower and softer". Lavinia asserted:

> There has got to be sincerity and a caring with your posing and your probing. You have to do that because otherwise it's interrogation. You have to be genuinely interested and soft with them because they are going to take a chance and expose possibly some personal stuff.

Lavinia's concern with students' feelings of vulnerability as it related to posing was a theme in her responses to the P-P-P Assessment. Her response to item 15 in the pre-lesson enactment survey, "What do you find most challenging about posing questions?" addressed this issue. Lavinia's response was, "I think the most challenging thing about posing questions is wanting to challenge the kids enough to get them to probe deeper into solving problems, to look at the different possibilities, without discouraging,

embarrassing, or frustrating the students who might be struggling."

Lavinia considered Gerald one of these "struggling" students and her response to the VSR protocol revealed that Lavinia intentionally manipulated her body positioning and language, voice and expression to encourage him to open up and perform. "Posing and probing is a way to honor kids" Lavinia reflected during VSR while unpacking the video clip that featured her interaction with Gerald.

This "honoring" and respecting the learners while, at the same time, doing formative assessment, requires that teachers take action to express this honor and respect to their learners by the ways in which they they pose, pause, and probe. This is important to include in descriptions of a developmental continuum of teachers' practice of FA. It is critical to eliciting a wide—and equitable—range of responses upon which teachers can then make better-informed instructional decisions. Currently, however, the posing construct map does not adequately address this aspect of actions teachers take to "elicit a wide range of responses."

I conclude by locating Lavinia's reflection on posing at the multistructural (3) level, with the analytic note that future iterations of the posing construct map, and the items design and scoring guides for the P-P-P Assessment need to take the salient qualities of Lavinia's VSR response into account as they are revised.

Pausing practice. Overall, evidence of Lavinia's planning for, enacting, and reflecting on pausing located her practice at the multistructural (3) level, intentional, supported pausing. At this level of pausing practice, the notion of differentiated pausing closely tied to purpose and context is not of prime concern to respondents. Respondents

whose practice is at this level can define reasons for pausing. During lessons in their classrooms, pauses—including some pauses longer than three seconds—occur and reflect intentional teacher actions to provide and protect wait time.

Planning. Lavinia's response to the lesson planning template evidenced planning for pausing. In the "Vocabulary Intro" component of her lesson plan Lavinia wrote "Think Time" in conjunction with the "Essential Question" for the lesson. This was interpreted as evidence of multistructural (3) planning for pausing, in which teachers plan pausing moves to elicit better quality responses from students though they may not be well-articulated.

On responses to items in the pre-lesson enactment survey, Lavina reported that she "often" planned pausing routines before lessons. However, in response to a Likert-scaled item with the statement, "When students need 'wait-time', I create specific routines for them" and the choices of Never, Rarely, Often, and Always as options, Lavinia chose "Rarely." This aligned with Lavinia's response to the lesson planning template regarding pausing. No evidence of planning "specific", tailored, or personalized pausing routines for students appeared. Presence of this would have indicated relational (4) or extended abstract (5) planning for pausing.

Enactment. Video-related enactment evidence (video records of the lesson and transcripts of the audio of the video) indicated multistructural (3), intentional, supported pausing. Lavinia's lesson featured "pair-shares" and "table talk" as "go to" pausing moves. Lavinia orchestrated these intentional routines for non-silent pausing during whole class and small group configurations. For example, and indicative of

multistructural (3) level pausing enactment, when Lavinia saw only three hands raised in response to a question she posed to the whole class toward the close of the lesson, Lavina inserted a "table talk" pausing move. Lavinia directed students, "Put your heads together, I'm going to call on any one of you. Talk first." Then, after pausing for a full three seconds, Lavinia primed their table talk with a suggestion, a verbal scaffold: "You may want to comment on what you saw from the other teams" (the students had just returned from doing a Gallery Walk about the room viewing and talking about other teams' work).

Analysis of the video evidence of Lavinia's pausing moves in the contexts of small group and whole class configuration revealed consistent and careful, intentional use of "wait time." One of Lavinia's consistently used pausing moves between her poses to an individual student or to students working in a small group was to acknowledge the feelings she observed as signs of being present for her students. For example, between her questions to one group of students Lavinia said, very slowly and gently, "It may cause a kind of frustration." This careful, intentional use of "wait time" indicates multistructural (3) level enactment of the pausing dimension of FA practice.

Reflection. The evidence of Lavinia's reflection on pausing generated by her engagement with the VSR protocol mirrored the qualities of her reflection on posing. Lavinia focused on the craft aspects of her practice as she unpacked video clips of her lesson enactment through the lens of pausing. Transcripts of the VSR session illustrate that Lavinia described what, how and why she paused (e.g., "to set up students like Bernetha, who has severe learning disabilities, to shine—and she did!"). However, this focus does not align well with the current version of the pausing construct map, which

describes respondents located at the multistructural (3) level as "able to reflect on motivations for increasing flexibility and strategic use of pausing moves (and accompanying scaffolds for) and offer suggestions on how." This description is "next steps"-focused. Lavinia, however, did not offer any next steps during VSR.

I resist locating Lavinia's reflection on pausing at the unistructural (2) level since the description of that level on the pausing construct map presumes the respondent has enacted pausing during the lesson(s) much less competently than Lavinia did. The description of reflection on pausing at unistructural (2) level reads:

> [Respondents] are able to reflect on specific opportunities for pausing that they missed and offer next steps to try to improve their pausing. They are able to reflect on why pausing is important for students [even if they did not orchestrate it skillfully or consistently in the lesson].

Therefore, even though Lavinia's reflection on posing is not well-aligned with the description of multistructural (3) level pausing in the current version of the posing construct map, I locate this facet of her pausing practice there. Revisions to the pausing construct map and the items design and scoring guides for the P-P-P Assessment should be considered in light of Lavinia's VSR response. It is empirical evidence that informs definition of the construct.

Probing practice. Overall, evidence generated by Lavinia's engagement with the P-P-P Assessment indicated that Lavinia's planning for, enacting, and reflecting on probing located her practice at the unistructural (2) level, task-focused probing. Respondents exhibiting unistructural level, task-focused probing contend the main purposes of probing are to spur student action and to make learners' thinking more visible, though their

actions may imply reasons and purposes for probing beyond just those. But these reasons or purposes are not expressed explicitly in planning, enacting, or reflecting. Responses to items/tasks indicate probing practice that relies on generic probing moves such as "Why?" or "What do you mean?" as "go to" probes that are beyond "probing to manage" or "probing to engage." At the unistructural (2) level of practice, probes may or may not elicit new information from learners.

Planning. Analysis of Lavinia's response to the lesson planning template suggests locating Lavinia's practice of this facet of probing at the unistructural (2) level. At the unistructural level, respondents plan specific and/or generic probes (e.g., "Why?" "How do you know?"), and often include their "go to" probes. Lavinia did this. She wrote verbatim probes in two components of the lesson: the "Application to the Real World" component (as students worked in groups on a "thinking map") and the "Further Tile Work" component as students were to "work on extending their patterns with tiles and through that try to ascertain what the rule might be." The probes were: "How do you know? Can you explain it to me? Hmm. I'm not sure I understand this pattern part you made. Can you help me? Do you mean...?"

According to Lavinia's pre-lesson enactment survey responses, completed before Lavinia began her lesson planning template response, some of the probes Lavinia planned in her lesson were her "go to" probes. In her response to survey item 29, Lavinia listed this probe: "I don't quite understand how you came up with your answer. Can you help me and explain what you did here?" as a probing question or statement that she found herself using again and again (a "go to" probe) with her students. Lavinia's responses to

the pre-lesson enactment survey and lesson planning template triangulate with the planning-related description of probing practice at the unistructural (2) level to support locating Lavinia's planning for probing at the unistructural level.

Enactment. Analysis of video evidence of lesson enactment suggests locating the enactment facet of Lavinia's probing practice at the unistructural (2) level. A key characteristic of probing at this level is that respondents tend to enact most of their probing in one component or portion of the lesson. Though Lavinia did probe an individual student's initial response during a whole class context in the closing component of the lesson (e.g., Lavinia probed Bernetha's one-word initial response of "Music" with "How come?"), almost all of Lavinia's probing occurred only during team work time in the "Application to the Real World" component of the lesson.

Consonant with the description of unistructural (2) level probing, live observation by the researcher, video recording of the lesson, and transcripts of the video recording revealed that some probes "worked" and made some learners' present thinking visible. There was also some evidence that what was elicited via probing was used by Lavinia. For example, during an extended interaction with a small group of two boys and two girls, Lavinia used one boy's contention (Harold's) that he was "constant all the time," which she had elicited through her probing of the small group, to challenge him on this notion as she related it to the ideas of continuous graphs and discrete graphs. She did so with good humor, "Are you always predictable? Do you ever get moody?" her eyes twinkling.

"Last Friday," Harold replied, with a laugh, smiling. Lavinia continued to use this honest admission from Harold playfully but respectfully (Lavinia knew him well and they had a strong, warm relationship) and attempted to help connect the ideas of constancy and unpredictability to the lesson's objective about continuous and discrete graphing. After all five of them laughed together, Lavinia summarized and brought up a previous activity they had engaged in together: the graphing of a tree's growth. Lavina probed them about that and how it related to what they were working on in the present moment.

Although this was an episode of skillful probing, thorough analysis of the totality of the enactment evidence supports the interpretation that Lavinia's enactment of probing reflects level two, unistructural probing, and not multistructural (level 3) probing. Lavinia's probing focused on spurring student action, making student thinking more visible. Significantly, Lavinia's enactment evidence did not reveal her encouragement or scaffolding of student-to-student probing, an important distinguishing feature often present in multistructural (3) enactment of probing. Multistructural (3) probing is more varied, features more consistent probing of "correct" answers, and seeks to aid in diagnoses and characterizations of student thinking by targeting misconceptions, than the probing evidenced in Lavinia's lesson enactment.

Reflection. The evidence of Lavinia's reflection on probing generated by her engagement with the VSR protocol mirrored the qualities of her reflections on posing and pausing. Lavinia focused on the craft aspects of her practice as she unpacked video clips of her lesson enactment through the lens of probing, In addition, she mentioned some of her purposes. But the descriptions of the current version of the probing construct map that

are overtly specific to the reflection facet do not address the purposes of probing directly. The lead-off descriptions of respondents at levels 2, 3, 4, and 5, however, do speak to the qualitative differences in the purposes behind respondents' probing moves. This allows me to locate Lavinia's reflection on posing at the multistructural (3) level and not the unistructural (2) level reflection on probing. The lead-off descriptions at these two levels read:

> *Respondents who* distinguish between probing to assist teacher decision making and probing to benefit student(s) being probed or any witness learner(s) witnessing the teacher-student(s) probing interaction (multistructural, (3) *targeted* probing).
>
> *Respondents who* contend the main purposes of probing are to spur student action and to make learners' thinking more visible, though their actions may imply reasons/purposes for probing beyond those. But these reasons/purposes are not expressed in planning, enacting or reflecting (unistructural, (2) *task-focused* probing).

Since Lavinia did express her purposes for her probing in her reflection, I locate her practice of reflection on probing at the multistructural (3) level. This excerpt from the transcript of Lavinia's response to the VSR protocol corroborates that location:

> Lavinia: With their laughter and also seeing how they fit that they could identify with this continuous graphing. It was making it where they could understand better. And then I pulled it back in [after our laughter together]. I pulled it back in and because I knew that – I knew they understood from their own – with their own selves, with their own hearts and minds they understood. It wasn't just this abstract thing. Now we are talking about their lives and their personalities and this is math beyond, this is me, who I am. Because I don't think that's going to be a lesson they will forget readily.

Areas for growth. Planning for probing and reflecting on posing stand out as areas for growth. Focused attention on these two areas in particular is likely to bring about

changes in practice that will be noticeable when analyzing practice along the continua articulated in the construct maps and scoring guides for the P-P-P Assessment.

Posing. Engaging in thinking about how questions posed might serve to inform instructional decision making after responses to the question are elicited, tagged, and considered is recommended. This is to develop the notion of contingent instructional moves in FA practice. Solid practice of relational posing demonstrates evidence that activities and pacing clearly reflect teacher decisions that are contingent upon student responses to questions posed about the learning target. Through such reflection, teachers often experience insights about how the FA moves, and when used in thoughtful, skillful combinations, can synergistically support their purposes.

Pausing. Lavinia's pausing was the most adept, consistent, and skillful of those observed in the study. Lavinia enacted a kind of differentiation with her pausing moves concerning individual students during group work that was not reflected in her planning evidence. An area for growth for Lavinia is for her to consider how she might differentiate her pausing moves even more, especially within the contexts of whole class configuration.

Probing. Planning families of probes designed to serve a greater variety of purposes (beyond getting students to elaborate and explain their work processes) is an area for growth and a recommended area to focus next step efforts. Developing knowledge of common misconceptions students hold about content will help in designing and deciding which probes to use for diagnosing misconceptions. More skillful probing cannot be developed without also developing deeper content knowledge and more sophisticated

pedagogical content knowledge in mathematics. Since no evidence of encouragement of student-to-student probing was observed in the course of Lavinia's engagement with the P-P-P Assessment, taking action to expect and scaffold student-to-student probing during class is also recommended.

Planning. Writing out potential probes, especially diagnostic-type probes, ahead of the lesson, and planning instructional pathways linked to likely responses to the probing, is recommended as a next step. Attention to p-prims and misconceptions that students are known to hold about the topic that is the target of the learning can help.

Enacting. Lavinia is already strong at leveraging her strong rapport with and knowledge of students to get them to engage with the content. An area for growth is for Lavinia to consider how to, and to try to, intentionally use the FA moves synergistically to support students' deeper cognitive engagement with the content. Additionally, an area for growth is that students could be taking more active roles in enacting FA moves.

Reflecting. Reflection on practice generated from engagement with the P-P-P Assessment was largely focused on craft aspects of posing, pausing, and probing, and purposes. An area for growth is to reflect on evidence of practice in ways that are anchored solidly in student thinking, rather than only focused on student engagement and affect. Also, during reflection with others, a focus on generating a number of alternative instructional decisions and pathways that could have resulted from the evidence that was actually elicited in the lesson—and from evidence that likely could have been elicited by employing a different data collection strategy (including trying a different question to pose)—is recommended.

Jessica

Background and context. Though Jessica was a second year teacher at the time of the study, and still attending Beginning Teacher Support and Assessment meetings, she had deep roots in the Kimm community. A native speaker of Spanish, Jessica had extensive experience tutoring and coaching local students. Jessica knew the siblings and parents of—as well as the students themselves—of many of her eighth grade math students long before she welcomed them across the threshold of her classroom.

Jessica holds a multiple subject credential. She had majored in childhood development, at the same local state institution where she had completed her credential program and the PACT. Completing her PACT portfolio was Jessica's only other experience reflecting on video clips of her instruction.

This was Jessica's second year teaching with CPM. Jessica had attended all the district-sponsored trainings on the CPM curriculum since she had begun teaching full time at Kimm nearly two years prior. "I like having that math community," Jessica reported. As an 8th grade math teacher, I also like learning about the expectations of high school math too." Her instruction was evolving. Jessica reported that this year she was focused more on the word problems and "getting students to explain their processes and show their work, show their steps and explain. Because that is what the tests ask for and when you go to jobs you have to be able to show them and also in words explain." Jessica identified the technology portion of CPM as one of its strengths, "because I can show the mini-lesson here and the kids can practice at home, online."

Jessica had 28 students in her block period "mathematics only" class that met three

days each week for 90 minutes. Demographics reflected the demographics of Kimm's student population (see Aaron's profile for percentages). Thirteen students in the focal class were designated ELs. One attended Special Day Class part-time.

The lesson, "Solving Equations with Fractions" occurred during a unit on systems of equations. Jessica's lesson plan conveyed that "38% of her students struggled with fractions, about 50% had mastered it", and 12% were "intermediate". At the time of the lesson, the class was "in the process of adopting a method to eliminate the fractions in systems of equations by using "fraction busters", lowest common multiple (LCM), or another option they have created."

The SLO was "SWBAT extend what they learned about solving equations with integer coefficients to equations that involve fractions and decimals. They will learn how to change fractional and decimal coefficients and constants to integers."

Jessica was targeting CCSSM 8.EE.7b: Solve linear equations with rational number coefficients, including equations whose solutions require expanding expressions using the distributive property and collecting like terms. The ELD Standard the lesson addressed was "Students will be able to describe both orally and in writing their process of solving equations with integer coefficients to equations that involve fractions and decimals."

Posing practice. Overall, evidence of Jessica's planning for, enacting, and reflecting on posing located her practice at the multistructural (3) level, constrained posing. As previously mentioned, respondents exhibiting multistructural (3) level constrained posing seem to be trying to get students to say what they, as the teacher, are thinking rather than eliciting from students a range of responses, including unknown responses and responses

surprising to the teacher. Planned questions connect to the learning target. Most or even all questions posed, however, are fact recall—what, when, where—and lower-level questions according to taxonomies such as Webb's, Bloom's, and Costa's.

Planning. The lesson planning template response evidenced that Jessica planned questions she considered checks for understanding of the lesson's objective, which is indicative of multistructural (3) planning for posing. In Jessica's case, each check was a math problem (e.g., in the "Assessment" section of the template, Jessica wrote: "Students will turn in an exit ticket that involves a problem similar to focus problem 5.16 that allows students to change fractional and decimal coefficients and constants to integers and they will do this by showing their work and explaining their process." This was directly tied to the SLO.

Jessica also wrote in the "Direct Instruction" component of her lesson plan "Check on work to see if they can adopt a way to solve this problem." Though verbatim questions have not been written out, the purpose behind any questions Jessica may ask while carrying out this "assessment strategy" —to check for understanding of one of the lesson's objectives—is clear and aligned with level 3, multistructural planning for posing.

Enactment. Enactment evidence indicated that Jessica's posing practice was located at the multistructural level (3), constrained posing. Poses during whole class configuration were uniformly closed ended. Students were expected to provide the one response the teacher had in mind, even when doing so appeared to be more of a hindrance than a help regarding students' advancement toward the learning target. Jessica sometimes used "fill-in-the-blank" question stems when formulating poses of this type,

such as “A decimal has...pause...what?” and “But you have to seek the growth by using...pause..what?”

Consonant with multistructural (3) posing, observation of enactment showed that during the lesson opening Jessica’s questions sought to elicit students’ prior knowledge (e.g., “Where have you heard this word constant before?” and “You have seen this before, what do we do to—talk to your group—how do you solve proportion step by step?”). Jessica’s posing during small group configuration featured questions about students’ processes for doing the work assigned, such as “What’s your process that you’re adopting?” and “How did you guys do it?” Yet a high percentage of poses were what/when/where, fact recall, and lower-level questions according to Bloom’s taxonomy, a core characteristics of multistructural (3) posing.

Reflection. Jessica’s reflection during VSR critiqued her posing. While she did not generate, during VSR, the actual wording of alternate poses she could have used, Jessica’s reflection included specific suggestions for why it was not a good pose. For example, Jessica said

> The way I stated [the question] was not well structured and it was about a topic [fractions] that only 50% of them have really mastered. A question not structured correctly about a concept they don’t understand—it was confusing to them. When I see that happening sometimes I’ll bring them back with, “So let’s summarize…”

Jessica’s responses to the VSR located her reflection on posing at the multistructural level (3).

Pausing practice. Each facet of Jessica’s pausing practice was located at a different level. Planning for pausing evidence indicated unistructural level (2), pausing enactment

evidence was located the multistructural level (3), and reflection on pausing evidence indicated location at the pre-structural (1) level. Overall, however, Jessica's practice of formative assessment encompassed by the pausing dimension as it is articulated in this study can arguably said to be at the unistructural (2) level. At this level, respondents want to avoid stretches of silence. Evidence generated from the pre-lesson enactment survey, the intake/planning interview, and the VSR protocol triangulate to support that Jessica's pausing practice, overall, resonates with this description and locates her pausing practice generally at the unistructural (2) level—unintentional, unsupported pausing—articulated on the pausing construct map.

As previously mentioned, respondents at the unistructural (2) level of pausing practice want to avoid stretches of silence and may feel fear, distrust, discomfort, or pain about silence in class, especially in response to questions they have posed or directions they have given. Responses to items/tasks indicate that silent pauses that do begin spontaneously are ended prematurely by teacher action. Student time on task or classroom discipline may confound with the pausing construct at this level of practice, particularly the aspect of the pausing construct related to silent pausing (versus non-silent pausing).

Planning. Jessica's response to the lesson planning template did not show evidence of planning for pausing. On responses to items in the pre-lesson enactment survey Jessica reported that she "often" planned pausing routines before lessons. She reported that "Often" "When students need 'wait-time', I create specific routines for them." Jessica

listed two pausing routines that she found herself using again and again: (a) "Take the time to think for specific time given", and (b) "First think, share, then write."

During the intake/planning interview Jessica reported that to her pausing is giving students specific directions on what to do. She explained

> If I tell students, 'Okay, take the time to think about this problem and analyze it,' automatically they go on to socializing. There has to be something more specific next to that. I really want to give them time to sit and think, but somehow it's not happening.

Together this evidence suggests that Jessica's planning for pausing is located at the unistructural level (2), *unintentional, unsupported* pausing. At this level, respondents plan lessons with few or no explicit pausing moves or routines.

Enactment. Enactment evidence indicated multistructural (3), intentional, supported pausing. Jessica orchestrated intentional routines for non-silent pausing during whole class and small group configurations: (a) by requesting students to work something out on their own white board, (b) instructing students to "talk in your groups" and (c) directing a pair of students to "talk about it [non-silent pause] and then do it." Indicative of multistructural (3) pausing enactment, when Jessica saw "not enough hands up" in response to one of her questions, occasionally she inserted a "pair share" or a "talk at your tables" move as a "go to" way to support pausing to increase student participation. Jessica did this three times during whole class configuration.

Reflection. Jessica's reflection on pausing located this facet of her practice at the prestructural level (1), pre-pausing. At this level, respondents are able to reflect on why it might be necessary to change their practices related to pausing, yet they are not

identifying specific opportunities for pausing that they missed and offering next steps to try to improve their pausing, as do respondents at the unistructural level of reflection on pausing. For example, during VSR, Jessica said:

> Pausing is not only a learning process for me but it's for them too because they're teenagers. They want to talk right away, they want to jump into the talking...
> This is only my second year and I've always felt like I'm still not – I'm still in the process of improving because it's a hard thing to do to pause, to give them that adequate time. And what's the adequate time? That's always in question.

During VSR, Jessica reflected that much of her pausing was "just within group discussion because that's what [students are] most comfortable with." Identification of this pattern of her pausing enactment during reflection falls short of identifying specific missed opportunities indicative of unistructural pausing. This reflection aligned with Jessica's response to pre-lesson enactment survey item 43, "What is the most challenging thing about pausing?" Jessica wrote, "Students may not feel comfortable with the quiet time. Students do not know how to use the time to think."

Probing practice. Overall, evidence of Jessica's planning for, enacting, and reflecting on probing located her practice at the unistructural (2) level, task-focused probing. As previously mentioned, respondents exhibiting unistructural (2) level, task-focused probing contend the main purposes of probing are to spur student action and to make learners' thinking more visible, though their actions may imply reasons and purposes for probing beyond just those. But these reasons or purposes are not expressed explicitly in planning, enacting, or reflecting. Responses to items/tasks indicate probing practice relies on generic probing moves such as "Why?" or "What do you mean?" as "go to" probes that are beyond "probing to manage" or "probing to engage." At the

unistructural (2) level of practice, probes may or may not elicit new information from learners.

Planning. Jessica's response to the lesson planning template indicates that she did not generate probes related to the learning target when planning. Therefore, her planning for probing is located at the prestructural level (1), pre-probing. In the pre-lesson enactment survey Jessica reported that she "often" planned her probing routines before lessons.

Enactment. Enactment evidence indicated unistructural, (2) task-focused probing. Indicative of this level of practice, most of the probing enacted in the lesson occurred in one component or portion of the lesson. For Jessica this was during the "Small Group" lesson component (I refer to the lesson component as Jessica did on her lesson plan/her response to the lesson planning template.) Probing enactment at the unistructural (2) level may or may not elicit new information from learners. During enactment, Jessica's probing did elicit new information from learners. Further, consonant with unistructural (2) probing, there was some evidence that what was elicited via probing was used by the teacher or a student or students. For example, after Jessica probed Tomás initial response in the lesson's opening (e.g., "How, Tomás? How is it growing?"), Jessica worked his response into her whole class conversation about a minute later, as she set up a whole class pose:

> Tomás said it shows an increase, from lowest to highest right, okay something that just keeps repeating, and we show it by using a number, the same number, so if I say that the tree, that when I bought it was five feet tall, and every year it grew four feet, listen...four feet every year what's the constant there?

Indicative of unistructural (2) probing, Jessica enacted probes tied to the learning

target. Most frequently her probes sought students to explain "how they changed fractional and decimal coefficients and constants to integers", one of two SLOs in her lesson plan.

Reflection. Jessica's responses to VSR included a few possible probes relevant to the learning target that were generic. Jessica's reflection on probing included her suggesting "and then that's when I would question [probe] 'Why are you using this? What do you guys understand?'" Jessica's reflective analysis of her probing practice described a "pattern" she carries out while interacting with small groups:

> What I'm doing is giving him time to explain so when I hear...so this is where I'm listening to what he knows and if he knows something that he can share with his groupmates. Then that's when I probe. So I first let one kid lead and then I kind of like dissect it by asking probing questions, that's what I do – that's usually the pattern I carry when I do that.

Areas for growth. Relative to other second year teachers whose practices of formative assessment I have observed, Jessica's enactment is strong in that there is evidence that Jessica is taking up what students are saying in response to her queries and attempting to use it productively to advance the learning of other students. This tendency will likely feed and multiply the effect of growth Jessica makes in other areas and aspects of the practice of formative assessment. Planning for probing and reflecting on posing stand out as the areas for growth most recommended for focusing attention in order to improve formative assessment practice.

Posing. Eliciting a wider range of responses could be achieved by several methods. Analysis of the kinds, types, and mix of questions posed during lessons (part of reflecting on posing) could help with formulating and prioritizing new strategies to try to elicit a

wider range of responses. “Structure” (content and formulation) of questions posed, timing of poses, and repetition of key questions posed are critical. Other FA moves — particularly priming for pausing, pausing, priming for bouncing, bouncing, and tagging— have significant roles to play too in efforts to get questions posed to elicit a wider range of responses.

Pausing. Analysis of current practice reveals that a high percentage of questions posed during whole class configuration are closed-ended. Interactions between Jessica and her students speed up when she seems to be hunting for a student to say the word or phrase that is on her mind before moving forward. During such exchanges, orchestrated supports for pausing are absent. Though Jessica’s use of improvised non-silent pausing moves such as pair-shares and table talk time is relatively strong, pausing moves during whole class configurations remain an area for growth.

Probing. Planning families of probes designed to serve a greater variety of purposes (beyond getting students to elaborate and explain their work processes) is an area for growth and a recommended area to focus taking a next step. Research that reveals p-prims and misconceptions students are known to hold about the topic that is the target of the learning can help. Additionally, taking action to expect and scaffold student-to-student probing during class is recommended.

Planning. Rapid improvement of FA practice will likely result from writing out potential probes, especially diagnostic-type probes, ahead of the lesson, and planning instructional pathways linked to likely responses to the probing. Planning can also include the act of representing, for example, a wide range of solution methods for a rich

task ahead of time. If Jessica does this during planning, she need not depend entirely upon only the solutions methods her students might come up with. Two additional benefits of such planning may be that (a) Jessica's content knowledge deepens and (b) that Jessica will be better primed to acquire richer pedagogical content knowledge (PCK) as she enacts and then reflects on how her lesson plans interacted with students. Teachers' PCK is enhanced from noticing and reflecting on what of their lessons plan they kept "as is" as they enacted it and what they improvised in attempts to meet the student needs they noticed in the moment. Planning can enhance improvisation and reflection.

Enacting. Using the FA moves synergistically to support teacher purposes is an area for growth. Over time, broadening the number of intentions (e.g., to include the intention of eliciting a wider range of responses) in one's instruction will deepen FA practice. Jessica is already strong at leveraging her strong rapport with and knowledge of students to push her students to work during class and work on their academic growth on their own time outside of class too. As Jessica broadens the purposes behind her moves and gains clarity about which intention she is prioritizing during enactment, her improvised enactment of a variety of FA moves will grow more sophisticated.

Reflecting. Reflecting on what factors contribute to classroom teacher-student interactions characterized by a series of tightly constrained poses that reflect a stance of "someone needs to say aloud the response I have in mind before we can move forward" could benefit Jessica's posing practice. One of the effects may be to become better able to discern students' thinking. Learning to distinguish when one is working to activate

students' prior knowledge from when one has fallen into "a student needs to say what's on my mind" is important in the developmental trajectory of a teacher's FA practice. Reflecting on practice with others has a role to play. Focus on generating a number of alternative possible actions/moves that could have been taken and generating several explanations for student action(s)/non-action that are grounded in evidence is recommended.

Aaron

Background and context. Aaron had always taught mathematics and science during his 16 years of teaching kindergarteners, third-, sixth-, and seventh-graders. The study occurred during Aaron's fourth year teaching math and science to seventh graders at Kimm, a PK-8 school of 723 students. All of Aaron's teaching had been at schools whose student populations comprised high percentages of underprivileged students and English learners. Kimm represents the school with the lowest percentage of low SES students at which Aaron had taught: 82%.

Aaron holds a multiple subject credential. He majored in liberal arts. To deepen his mathematics knowledge, while teaching full-time Aaron had both taken and audited mathematics courses at a community college. He had served as a math coach for two years (in 2010-2011 and 2007-8). During the study, Aaron was pursuing his administrative credential.

Aaron had not experienced reflecting on video clips of his instruction before. He expressed dissatisfaction with evaluations of his teaching that told him, "Great job!" Aaron wanted both "better feedback" and "to know how I stack up." Aaron was active in

district-sponsored professional development related to math instruction, attending several meetings a year.

Aaron had been a lead teacher in the piloting of the College Preparatory Mathematics (CPM) curriculum he used. The study occurred in his fourth year teaching with CPM. Aaron appreciated that CPM "encourages students to think on their own," had "different jobs for students to do: facilitator, resource manager, reporter," and expected "students to work together."

Aaron chose to plan, enact, and reflect on a lesson for his seventh grade mathematics class that met three days each week for 90 minutes. The class was a "mathematics only" class that did not aim to integrate science.

Students in Aaron's class reflected the demographics of Kimm's student population: approximately 46% Hispanic or Latino, 43% Asian, 5% Filipino, 3% two or more races, 1% white and 1% Black or African American. Forty-seven percent of students school-wide were officially designated English learners (ELs). Of 27 seventh graders in Aaron's class—18 boys and nine girls, twelve were designated ELs (approximately 44%). Two had 504 plans. One of Aaron's students attended Special Day Class some of the time.

Aaron's 90-minute lesson with an "at level" class of 27 seventh graders—18 boys, 9 girls—occurred within a unit on probability in the College Preparatory Math (CPM) curriculum. Twelve of Aaron's students were officially designated English language learners.

The lesson occurred during a unit on probability and expected students to use a probability tree to model outcomes for compound events. Students were expected to have

used “the probability array” before and “understand that probability is based on understanding the number of outcomes over the number of trials.” Aaron’s response to the lesson planning template did not mention ELD standards. His response listed the lesson as addressing four CCSSMs about statistics and probability: standards 5, 6, 8a and b, all of which belong to the cluster “investigate chance processes and develop, use, and evaluate probability models.” Figure 14 presents these standards in full.

Posing practice. Overall, evidence of Aaron’s planning for, enacting, and reflecting on posing located his practice at the multistructural (3) level, constrained posing. As previously mentioned, respondents exhibiting multistructural (3) level, constrained posing seem to be trying to get students to say what they, as the teacher, are thinking rather than eliciting from students a range of responses, including unknown responses and responses surprising to the teacher. Planned questions connect to the learning target. Most or even all questions posed are fact recall—what, when, where—and lower-level questions according to taxonomies such as Webb’s, Bloom’s, and Costa’s.

Investigate chance processes and develop, use, and evaluate probability models. [cluster]

Standards:

5. Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0 indicates an unlikely event, a probability around 1/2 indicates an event that is neither unlikely nor likely, and a probability near 1 indicates a likely event.
6. Approximate the probability of a chance event by collecting data on the chance process that produces it and observing its long-run relative frequency, and predict the approximate relative frequency given the probability. *For example, when rolling a number cube 600 times, predict that a 3 or 6 would be rolled roughly 200 times, but probably not exactly 200 times.* (p. 50)

8a and b.

8. Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation.
    a. Understand that, just as with simple events, the probability of a compound event is the fraction of outcomes in the sample space for which the compound event occurs.
    b. Represent sample spaces for compound events using methods such as organized lists, tables and tree diagrams. For an event described in everyday language (e.g., "rolling double sixes"), identify the outcomes in the sample space which compose the event. (p. 51)

Figure 14. Grade 7 standards within the Statistics and Probability domain of the California Common Core State Standards-Mathematics Aaron identified his lesson as addressing.

Planning. Aaron did not respond to the "Lesson Implementation" section of the lesson planning template. No inferences can be drawn from an item that elicits no

response. Aaron's responses to the "Lesson Overview" and "Teacher Preparation" sections of the template did not list any questions planned as checks for understanding or otherwise. Multistructural (3) planning for probing includes planning questions as checks for understanding.

Aaron reported during the planning-focused interview that he did not lesson plan for student questions, a response all study participants shared. Aaron did not take the pre-lesson enactment survey, which contained items that targeted planning for posing. Due to a lack of response, Aaron's planning for posing cannot be soundly located on the continuum of practice.

Enactment. Aaron's posing enactment was located at the multistructural (3) level. Indicative of multistructural posing, Aaron's questioning succeeded in eliciting students' prior knowledge related to the learning target, probability. His posing during whole class configuration, however, fell into "guess what the teacher is thinking exchanges" and frequently conformed to the I-R-E (teacher inquiry, followed by student response, followed by teacher evaluation) discourse pattern (Mehan, 1979). Though Aaron did pose a mix of questions, overall, a high percentage (over 80%) of questions posed were lower-level according to Bloom's taxonomy or were closed-ended questions. During whole class instruction, Aaron's posing seldom elicited a wide range of responses, which is a characteristic of relational (4) level posing practice. A representative exchange from the opening of Aaron's lesson:

> Aaron: Probability is based on what? Based on... what, what does our answer have to be? *One-second pause.* When we are done with probability, when we're trying to figure out the chance of something happening? It has to be what? *Three seconds of silence. No hands raised.* Talk to

your neighbor. What does probability have to be at the end?
Elias: Percent.
Aaron: Good. It's going to be a percent of something happening.

Aaron's posing during small group configuration demonstrated a greater mix of questions than during whole class configuration, but not consistently enough to indicate his location in posing at the relation (4) level, *flexible* posing. In one especially strong episode of interaction with a group of four students, Aaron asked 11 questions: two about students' work processes generally, three about specific math content/procedures, three that sought information about students' understanding, and two focused on encouraging student-to-student interactions about math content/processes within the group (e.g., he asked June, "Why don't you ask Ly what she is doing?").

Reflection. Aaron's reflection during VSR included specific suggestions for how to improve posing, indicating multistructural reflecting on posing. Reviewing a video clip of a small group work Aaron said, "See, but I didn't ask her to prove it. I should have asked her to prove it." Identifying this missed opportunity for a question to be posed as a check for understanding indicates multistructural (3) level reflecting on posing.

Pausing practice. Overall, evidence of Aaron's planning for, enacting, and reflecting on pausing located his practice at the multistructural (3) level, intentional, supported pausing. As previously mentioned, at this level of pausing practice, the notion of differentiated pausing closely tied to purpose and context is not of primary concern to respondents. Respondents can, and do, explain reasons for pausing. During lessons, teachers at the multistructural (3) level of practice take intentional actions to provide and

protect wait time. Pauses longer than three seconds long occur related to procedures that teachers at the multistructural (3) level or pausing practice set up and carry out.

Planning. Aaron did not respond to the pre-lesson enactment survey. Items targeting planning for pausing in that survey therefore could not provide evidence used to help locate this facet of Aaron's pausing practice. Aaron's incomplete response to the lesson planning template provided no evidence of planning for pausing.

Enactment. Enactment evidence indicated multistructural (3), intentional, supported pausing. Aaron's lesson enactment evidence included "pair-shares" and "table talk" as "go to" or consistently used pausing moves. He demonstrated verbal and nonverbal support of pausing, such as by saying, "Hang on, hang on!" to impatient students wanting to talk and by signaling through hand motions "stop" and "easy, slow down" at other times when students were starting to speak before the pause seemed sufficient to Aaron. Aaron's lesson included several whole class silences that lasted from a few to several seconds long (3-12 seconds). At times, when no hands were immediately raised after Aaron had posed a question the whole class, Aaron "threw in a pair-share", as he referred to his actions while reflecting on his video clips during the reflection session interview. During VSR, Aaron explained, "When I see blank looks, I'll throw in a pair share like that." Aaron used other moves to support pauses too. For example, evidence from the lesson transcript reveals that when a student was faltering in the spotlight, Aaron re-primed, encouraging her, "Any answer is fine. Pick your brain. You gotta think yourself through this."

Another kind of intentional pausing Aaron enacted was "walking away" and "checking back" with a student. During VSR Aaron said, "That's an intentional pause. I don't want to just sit there. What I'm hoping for is that when I pause [like that, by saying "I'll check back"] if they are working in a group they might try to pick the brain of somebody else."

Reflection. Aaron's reflection on pausing located this facet of his practice at the unistructural (2) level, which is characterized by, in part, respondents being able to reflect on why pausing is important for students. During VSR, Aaron explained about one of the silent pauses during whole class instruction that lasted several seconds, "Everybody has got to be patient. It's respecting somebody's voice. If you ask a question, you need to give them time to think. That person needs time to think. I am patient." As Aaron continued reflecting, his response suggested that though he did not consistently enact relational (4) level pausing during enactment, he knew how important it was to approach pausing flexibly. His words suggested that he believed that pausing approaches taken need to intentionally relate to the student or students involved, and that pausing needs to be differentiated for and tailored to students, a quality of pausing at the relational (4) level of practice. About whole class silences that stretched out while one student was "on the hook" or "in the spotlight" and expected to answer, Aaron said during the reflection session interview: I go too far in that sometimes. I will just wait and wait and wait, and that's kind of putting pressure on them. You have to know the kid."

Probing practice. Overall, evidence of Aaron's planning for, enacting, and reflecting on probing located his practice at the multistructural (3) level, targeted probing. As

previously mentioned, respondents exhibiting targeted probing distinguish between probing to assist teacher decision making and probing to benefit student(s) being probes or any witness learner(s) to the teacher-student(s) probing interaction. Responses to items/tasks indicate probing is potentially valuable to teachers or student decision making. Probes target uncovering misconceptions. At the multistructural (3) level of probing practice, teachers leverage probes and what probes make visible.

Planning. The only responses related to planning for probing that Aaron provided were his responses to questions during the planning interview (SSIP-1). Aaron reported he does not explicitly plan for probing. By self-report only, therefore, Aaron's planning for probing is located at the unistructural (2) level. At this level of practice, respondents do not generate probes related to the learning target when planning.

Enactment. Enactment evidence indicated multistructural (3) probing. Indicative of this level of practice, Aaron probed correct answers. Aaron did this most frequently during group work, such as when he asked Ly, as Ly was showing Ramon how to set up the problem, "Yeah, because why?" Aaron probed students' responses in every component of the lesson (launch, explore, discuss and summarize) and in every configuration (one-on-one, small group, whole class). During the lesson, Aaron consistently encouraged student-to-student probing, such as when he asked June, "Why don't you ask Ly what she is doing?" Both are characteristic qualities and actions of multistructural (3) level probing practice.

Reflection. Aaron's reflection on probing demonstrated that he was able to conjecture on possible alternate post-probe pathways for instruction, a distinguishing quality of

multistructural (3) reflecting on probing. For example, during VSR, he identified an opportunity during whole class instruction where he could have inserted a probe to all: "I should have said, "Of, what is that? When you say of, what does that mean?" Aaron continued, "That would have been cool, because then I could have got back, "Multiplication," and "Oh, that's how you calculate it!"

Aaron also spoke to how improved probing might have improved options for pathways for instruction, another indication of multistructural (3) reflecting on probing. Regarding the same scenario, Aaron further explained:

> That would have been perfect if I would have said that. If I did this all over, I would have been, "Of!" underlined it, [saying exactly what he would say to the class in this alternative version of the lesson] "Oh, we're going to keep that word." (Time passes in Aaron's imagined scenario…) "Now we're going to come back to it. Oh, let's do some probability!" And then calculate it.

Areas for growth. Overall, Aaron's formative assessment practice would benefit from (a) a focus on planning for pausing, especially whole class pauses before any bouncing occurs, and (b) planning for probing. Anticipating potential student confusions, sticking points, and misconceptions could be a helpful step toward generating content-specific probes ahead of time. Given his long experience teaching mathematics, Aaron should have little trouble with this first step.

Posing. Taking action to enact posing that elicits a wider range of responses (e.g., through priming and pausing moves tied to poses) is an area for growth. Given that students' questions are helpful in determining their zone of proximal development in a domain, planning for student questions—and supporting students to ask questions at

several key moments in the lesson is recommended. Current practice suggests no regularly enacted routines that foster students asking questions.

Pausing. Focusing attention on pausing moves/routines before any bouncing moves are made should serve to widen the response space and increase the variety of voices that get heard during classroom discourse. This is an area for growth. Students can be recruited to help with protecting and monitoring this kind of pausing. Inviting students' reflection and analysis on benefits they perceive from the enactment of this kind of pausing can be especially motivating to a community of learners, encouraging them to continue it.

Probing. Planning families of probes designed to uncover misconceptions and confusions is an area of growth worth pursuing. Some curriculum is better than others in communicating to teachers areas where students typically get stuck, revealing p-prims they often hold, and identifying and explaining misconceptions students are known to subscribe to. It is recommended that as Aaron pulls in curriculum from outside sources to supplement the district-adopted curriculum, he should critically analyze (a) the extent to which he anticipates the challenges and confusions his students may have with this curricular material and (b) the extent to which that curriculum and its instructor supports assist in determining these challenges and confusions for students and offer ideas—such as specific probes to use—for how teachers can uncover, diagnose, and productively address them.

Planning. Rapid improvement of FA practice will likely result from focusing on: (a) planning for whole class pausing in particular, (b) writing out potential probes (especially

diagnostic type probes) ahead of the lesson, and (c) planning instructional pathways tied to possible and likely responses to the probing. It is recommended that of these three areas for focus on planning, Aaron prioritize "a." Focusing on "a" will likely, almost inevitably, benefit the classroom community, improve the quality of class discourse, and serve to increase and widen the response space. These benefits will make efforts at focusing on "b" and "c" more fruitful. For example, planned whole class probes (resulting from focusing on "b") deployed without pausing the whole class before bouncing to a student would likely limit the power and utility of those probes. Because the FA moves inter-relate, Aaron's attention to their integration is essential to leveraging his use of the moves to advance student learning and optimize his instructional decision making.

Enacting. A priority area of growth is improving pausing enactment during whole class configuration in this context in particular: after a question has been posed and before any student has been called on to respond. A more sophisticated practice of pausing should serve to improve Aaron's bouncing, a practice critical to gaining a better quality snapshot of the class's thinking. The goal is more sound instructional decision making and increased attention to the number and percentage of students who participate in whole class discourse and other teacher-student interactions during the lesson. Aaron's own need for pauses—a cognition-based need that all teachers share—may also be supported by pre-bounce pausing routines that can help increase access to academic content for students' who may not be especially quick thinkers.

Reflecting. An area for growth during reflection concerns focusing on the quality of posing, with particular attention on types, kinds, and categories of questions. Reflecting on what factors contribute to classroom teacher-student interactions following the I-R-E pattern could benefit Aaron's posing.

Eliza

Background and context. Eliza was in her eighteenth year of teaching mathematics at the time of the study. Not all Eliza's teaching had been at Sierra Middle School. She held a multiple subject credential and a single subject credential in mathematics. Eliza had majored in mathematics, and had earned a master's degree in mathematics education. For the past five years she had been teaching mathematics at a local community college in addition to teaching middle school mathematics.

Eliza reported having been a member of the National Council of Teachers of Mathematics (NCTM) and having attended conferences sponsored by NCTM. Quarterly, Eliza had been attending professional development meetings on the teaching of mathematics. The professional development was sponsored by an organization dedicated to helping eight local school districts improve the mathematics instruction students in those districts received, especially in light of the demands of the new Common Core state standards in mathematics (CCSSM) and the new state tests accompanying them.

It was Eliza's second year using the curriculum she used with her focal class, an advanced eighth grade class, for the study: Mathematics Vision Project (MVP). MVP's curriculum modules aligned with the CCSSM. According to MVP, a teacher using MVP connects the Eight Mathematical Practices to content by "launch[ing] a rich task and then

through ‘teacher moves’ encourag[ing] students to explore, question, ponder, discuss their ideas and listen to the ideas of their classmates” (MVP, 2017).

Eliza chose to plan, enact, and reflect on a lesson for an advanced eighth grade mathematics class that met daily for 50 minutes. Students in Eliza’s class reflected the demographics of the population of Sierra Middle School, a school of 760 seventh and eighth graders: 56% Hispanic or Latino, 37% Asian, over 10% of whom spoke Vietnamese at home, 2% Filipino, 2% two or more races, and 1% white. Eighty-three percent were officially designated “socioeconomically disadvantaged.” Twenty-one percent were classified English learners. Five students in Eliza’s focal class of 30 students were English learners.

The lesson Eliza planned, enacted, and reflected on occurred during a unit on linear and exponential functions and was titled “Getting Down to Business.” The lesson expected students to compare the business plans of two imagined companies, Calc-u-rama and Computafest. Each was developing a plan for future growth. The SLO was “Solidify understanding of linear and exponential functions through a real life problem task.”

According to Eliza’s response to the lesson planning template, the lesson was targeting three CCSSM about functions and incorporating two mathematical practices:

> F-LE.2 Construct linear and exponential functions, including arithmetic and geometric sequences.
> F-BF.2 Write arithmetic and geometric sequences both recursively and explicit formula.
> F-IF.7 Graph functions expressed symbolically and show features of the graph.
>
> Mathematical practice 4: Model with mathematics.

Mathematical practice 5: Use appropriate tools strategically.

Eliza's response to the lesson planning template identified the lesson as addressing ELD standards that had students "Exchanging information and ideas with others through oral collaborative discussions on a range of social and academic topics," SL 8.1, 6; L. 8.3,6".

Posing practice. Overall, evidence of Eliza's planning for, enacting, and reflecting on posing generated from engagement with the P-P-P Assessment located her practice at the multistructural level (3), constrained posing. As previously mentioned, at this level respondents' posing seems to be trying to get students to say what they, as the teacher, are thinking rather than eliciting a range of responses from students, including unknown responses and responses surprising to the teacher. Questions planned connect to the learning target. However, most or even all questions posed, are fact recall—what, when, where—and lower-level questions according to taxonomies such as Webb's, Bloom's, and Costa's.

Planning. The lesson planning template response evidence supported locating Eliza's planning for posing at the multistructural (3) level. Characteristically at this level respondents will plan questions they consider checks for understanding of the lesson's objective. Eliza's response to the lesson planning template included plans to use these three questions as checks for understanding: (a) "What is the difference between linear and exponential functions? (b) Which company would you choose to invest in and why? (c) Is this function discrete or continuous?" The first question, planned for use at the end of the lesson, is the pose most closely aligned with the stated SLO.

Eliza's response to a Likert-scaled item on the pre-lesson enactment survey (item 12) revealed that by self-report Eliza "rarely" planned posing routines before her lessons. This self-reported evidence does not change the location of Eliza's planning for posing on the continuum of posing practice, but it does help to contextualize it. For Eliza, having written the questions she will use in the lesson ahead of time was a rarely occurring lesson planning activity.

Enactment. Live observation by the researcher, video recording of the lesson, and transcripts of the video recording served as evidence used to locate Eliza's enactment of posing. Analysis of enactment evidence supported locating her practice of this facet of posing at the multistructural (3) level, constrained posing. Analysis of the lesson transcript showed that Eliza posed a high percentage of low-level and closed-ended questions. Seventy-seven percent of all questions recorded were "dichotomous answer choice questions." In this type of question formulation, Eliza asked students to choose between two options, and provided the two options in the pose itself. One option was always the correct answer. Video and transcript evidence shows, for example, that Eliza asked: "The line goes on forever or stops at a certain number year?" "Should we connect it or shouldn't it be connected?" "Is it discrete or continuous?" Analysis of enactment evidence reveals that, consonant with multistructural (2) posing, questions posed seldom elicited a wide range of responses.

Reflection. Analysis of Eliza's response to the VSR protocol determined that her reflecting on posing should be located at the unistructural (2) level on the posing construct map. At this level, respondents are able to reflect on benefits that might accrue

from using a questioning scheme. Eliza posited that if she tried out a questioning system—one she dubbed "four corners and center" that she would benefit by: (a) receiving more student responses, (b) keeping more students engaged, and (c) helping her better determine which students understood the SLO and which did not—yet.

Pausing practice. Analysis of the body of evidence on Eliza's planning for, enacting, and reflecting on pausing located her practice at the unistructural level (2), unintentional, unsupported pausing. As previously mentioned, respondents at the unistructural level want to avoid stretches of silence and may feel fear, distrust, discomfort, or pain about silence in class, especially in response to questions they have posed or directions they have given. Responses to items/tasks indicate that pauses that do begin spontaneously are ended prematurely by teacher action that interrupts the pausing. At this level, teachers tend to miss cues that pausing is needed. With unintentional, unsupported pausing, student discipline or time on task can confound with the construct. Classroom action related to pausing is not well organized or modulated.

Planning. Analysis of Eliza's responses to the lesson planning template, the pre-lesson enactment survey, and the intake/planning session interview located Eliza's planning for pausing at the unistructural (2) level on the continuum of practice articulated in the pausing construct map. Eliza's response to the lesson planning template included no pausing moves or routines, indicative of planning for pausing at this level. Eliza's self-report on the Likert-scaled item about planning for pausing in the pre-lesson enactment survey (item 27) asserted Eliza "rarely" planned her pausing before lessons. During the intake/planning and reflection interview sessions, Eliza reported "I'm always rushing."

Eliza also shared in those interviews that she was planning to sing the A, B, C's in her head after posing a question in order to help "slow herself down" during the video recorded lesson. Eliza's responses to the survey, lesson planning template, and intake/planning and reflection session interviews support locating Eliza's planning for pausing at the unistructural (2) level.

Enactment. Live observation by the researcher, video recording of the lesson, transcripts of the video recorded lesson, and Eliza's response to the VSR protocol were used to locate Eliza's pausing enactment on the pausing construct map. Analysis of the body of enactment evidence supported locating Eliza's practice of this facet of pausing at the unistructural (2) level, unintentional, unsupported pausing. At this level, respondents tend to enact lessons and student-and-teacher "dialogues"/exchanges that reflect their discomfort with silence, especially with moments of whole class silence after they have posed a question. This was the case for Eliza. Analysis of video evidence revealed that Eliza's lesson enactment featured one silent pause that lasted longer than two seconds and that it occurred after Eliza posed a question to a small group. No pauses over two seconds long occurred during whole class instruction.

At the unistructural (2) level of pausing practice, respondents want to avoid stretches of silence. They may feel fear, distrust, discomfort, pain about silence in class, especially in response to questions they have posed or directions they have given. This description aligns with Eliza's descriptions of her perceptions of her experiences of trying to pause during the lesson. During VSR, Eliza spoke to how unnatural, challenging, and different pausing—and not talking—during class time felt to her. While unpacking a video clip,

Eliza said, “At this moment [on the video] I was practicing pausing, practicing not to talk. It was hard! It was not natural for me!” Analysis of the body of evidence about enacting pausing generated by Eliza’s engagement with the P-P-P Assessment supported locating this facet of her pausing practice at the unistructural (2) level, unintentional, unsupported pausing.

Reflection. Analysis of Eliza’s reflection on pausing generated by her engagement with the VSR protocol supports locating her skill in this facet of practice at the unistructural (2) level. According to the construct map for pausing, at the unistructural (2) level respondents are “able to reflect on specific opportunities that they missed and offer next steps to try to improve their pausing. They are able to reflect on why pausing is important for students.” The scoring guide for pausing notes that reflections at level two “include suggestions for including pauses in lessons, often tactics (e.g., “I will count to 5 in my head)” and “explores reasons for rushing.” Eliza’s responses to the VSR protocol illustrate these characteristics:

> Before I started the lesson, I kept saying to myself, “Sing the ABCs.” But my pauses looked so short! I did not pause. I thought I could ask a question and then be quiet, and not keep rephrasing the question again and again, assuming they didn’t understand it. Students need to process that question. We need to give them that chance.

In this response Eliza explores a possible reason for her “rushing”. Eliza conjectures that one reason she may be not pausing is that she assumes students “didn’t understand” the question she asked. Therefore, she keeps rephrasing the question rather than let silence hang in the air.

During VSR Eliza identified several specific opportunities where she could have

paused and suggests a tactic she will try in the future. One example of this occurred while Eliza was unpacking a video clip of whole class instruction:

> So I asked, "What is the net income?" I didn't wait for them to answer. I asked again, "What is the net income?" Then I clarified my question again. So I didn't give them a chance to think about what I had asked them. That's not good pausing in my opinion. So I am going to count.

Probing practice. Overall, evidence of Eliza's planning for, enacting, and reflecting on probing generated from engagement with the P-P-P Assessment located her practice at the unistructural (2) level, task-focused probing. As previously mentioned, respondents exhibiting unistructural (2) level, task-focused probing contend the main purposes of probing are to spur student action and to make learners' thinking more visible, though their actions may imply reasons and purposes for probing beyond just those. But these reasons or purposes are not expressed explicitly during planning, enacting, or reflecting. Responses to items/tasks indicate probing practice relies on generic probing moves such as "Why?" or "What do you mean?" as "go to" probes that are beyond "probing to manage" or "probing to engage." At the unistructural (2) level of practice, probes may or may not elicit new information from learners.

Planning. Analysis of Eliza's response to the lesson planning template and the pre-lesson enactment survey support locating her planning for probing at the unistructural (2) level. At this level, respondents plan specific and generic probes (e.g., "Why?" "How do you know?"), and often include their "go to" probes in their plans. Eliza's response to the lesson planning template revealed Eliza planned to ask the probe "Why?" twice during her lesson. Eliza's response to pre-lesson enactment survey item 29, which was "Please list any probing questions or probing statements (e.g., 'Can you explain?') that you find

yourself using again and again with your students, identified Why? as one of Eliza's "go to" probes. Why? is a fine, although generic, probe. No other verbatim and specific probes appear in Eliza's lesson plan. No routines to support probing were identified in Eliza's response to the lesson planning template.

Enactment. Analysis of enactment evidence (the video recording of the lesson and the transcription of the video recording) served to locate the enactment facet of Eliza's probing practice at the unistructural (2) level, task-focused probing, on the continuum of practice articulated by the probing construct map. Video evidence showed that Eliza relied on "How come?" as a probing move she used again and again, one of her "go to probes." This is illustrative of unistructural (2) level practice of probing. Unistructural probing is also characterized by the fact that most of the probing done by the respondent occurs in one component or portion of the lesson. Video evidence compared to Eliza's response to the lesson planning template confirmed that Eliza enacted all her probing during the "Explore" phase or component of her lesson as students worked in groups.

Teacher encouragement of student-to-student probing is not evident at the unistructural (2) level of probing practice, though it may be at the multistructural (3) level. In alignment with that expectation, video enactment evidence did not reveal any teacher encouragement of student-to-student probing.

Reflection. Analysis of Eliza's response to the VSR protocol located her practice of reflecting on probing at the unistructural (2) level. At this level, respondents are "able to reflect and suggest alternate probes that, while related to the learning target, still may be generic or task-focused only." During VSR, Eliza identified her repeated use of "How

come?" as a "simple probe." While unpacking a video clip of her interaction with a small group, Eliza suggested a specific/verbatim probe she could have used in place of her having asked How come? During VSR Eliza suggested "Can you give me another example of linear?" as an alternate probe she could have used.

Eliza's reflecting on probing is located at the unistructural (2) level. Its content, however, suggests that Eliza knows what she needs to do next to develop her probing skills, which is noteworthy. Responses generated during VSR point to Eliza's probing practice being on its way to becoming multistructural (3). During VSR, Eliza states that she "could develop that question [the alternate probe she just offered, i.e., "Can you give me another example of *linear*?"] into a more rich question." Eliza is not yet able, however, to come up with what that richer probe could be in that moment. Yet, significantly, while reflecting Eliza communicates some of the characteristics of a better probe. Eliza's description of better probes—and the ideal function thereof— aligns directly with characteristics of multistructural (3) probing.

Eliza said, regarding better probing with richer questions, "You have to rephrase what the student did or said. You could say, 'Can we think of an example of that?' So, I mean, I am not sure. I'm learning about that." During multistructural (3) probing, teachers incorporate students' words into probes as they attempt to leverage one student's response for the benefit of other students. These "other students" are students "witnessing" the probing, sometimes a desk partner, table mates, group members, or all the other classmates who are listening in. So while the probing dimension of Eliza's formative assessment practice is not yet at the multistructural (3) level, Eliza's skills in

the reflecting facet of probing have helped determine that Eliza is on her way to probing at this next level. Eliza knows where she needs to go with her probing.

Areas for growth. Areas for growth that stand out are planning for and enacting pausing moves and routines to support student think time and the elicitation of a wider range of responses students. Eliza's current pausing practice suggests that the voices of slower-to-respond students are not getting consistent opportunities to be heard in the question-and-answer exchanges that occur during whole class configuration.

Posing. Eliciting a wider range of responses could be achieved by several methods. Analysis of the kinds, types, formulation of and mix of questions posed during lessons could help her formulate and prioritize new strategies to try to elicit a wider range of responses. The affordances and constraints of using "dichotomous answer choice questions" warrant close attention since they are such a prevalent and salient characteristic of Eliza's posing. An area for growth is to work in other kinds and types of questions and question formulations (e.g., open-ended or metacognitively-focused questions or statement "poses" such as "I'm not sure I see what you mean.") Consideration of how other FA moves can synergistically support posing and students' engagement with question is recommended.

Pausing. An area for growth related to pausing is for Eliza to examine the orientation she holds toward incorrect answers. Some of Eliza's "rushing" through students' needed think time may be related to (a) Eliza assuming that students have not understood her question if her question is met by immediate, whole class silence and (b) Eliza's not wanting to receive incorrect answers as responses. Incorrect answers merit coaxing to be

elicited and deserve welcoming once they are. Some teachers have found that their pausing practices change for the better when they come to see that "wrong answers" will not spread like an epidemic through the classroom if the teacher tags a "wrong answer" on the board.

Probing. Eliza's reflection on probing outlined a fruitful next step to take to develop her probing: to practice rephrasing what a student has said and incorporating that into a probe. Another area for growth concerns planning for probing. Planning families of probes designed to serve a variety of purposes is recommended. Since no evidence of encouragement of student-to-student probing was observed in the course of Eliza's engagement with the P-P-P Assessment, taking action to expect and scaffold student-to-student probing during class is also recommended.

Planning. Planning for the pauses teachers need during lessons is essential. This is an area for growth for Eliza. Writing out potential probes, especially diagnostic-type probes, ahead of the lesson, and planning instructional pathways linked to likely student responses to the probing, is recommended as a next step. Priming moves specific to each probe or family of probes can be decided ahead of time too and is recommended.

Enacting. Students could be taking more active roles in the doing and supporting of FA moves during lessons. Focusing on scaffolding students' active roles in the moves, so that students are not just "recipients" of them, the ones whom the moves are enacted "upon," will support the enactment of more developmentally appropriate pedagogical practices for middle school students. In the middle grades, students are becoming ready to assert increasing autonomy and assume greater responsibility for their learning. Efforts

to align pedagogical practices with developmental changes in adolescent cognition and social behavior (Krajcik & Czerniak, 2007; Eccles, Midgley, & Adler, 1984; Eccles, Lord, & Midgley, 1991) —such as by fostering students' roles in initiating, carrying out, and analyzing the class's use of the FA moves—are likely to engage and support middle grade learners.

Reflecting. Eliza's reflection has already brought her ideas for next steps (i.e., access a "question bank") and places to focus on (i.e., try the "four corners and center" method). A suggested next step is to reflect on how to take action to bring some of those ideas to life. For example, reflection on practice has already helped Eliza to realize she needs to "wait more". Eliza can ask herself: What are five different ways I could try to support myself in "waiting more"?

Selena

Background and context. At the time of the study Selena had taught secondary school mathematics for nearly 30 years. Some of that teaching occurred abroad. Selena had been teaching mathematics at Sierra Middle School, a school of 760 seventh and eighth graders, for several years. She was presently serving as mathematics department chair. Selena had majored in mathematics and held a single subject credential.

Though not a current member, Selena had been an active member of the National Council of Teachers of Mathematics for over a decade, participating in annual conferences and networking with other math teachers. When the Common Core State Standards in Mathematics came along, Selena became the primary middle school teacher from her district to attend informational meetings at the county office of education.

(Selena's district, the district in which all the teachers in the study taught, has 13 elementary schools and three middle schools.)

About her participation, Selena reported, "At the time, almost everybody disliked Common Core. Whereas I thought, 'No, this is actually the way to go. Students should learn the Common Core way. Let's get out of that giving worksheet mode!'"

It was Selena's fourth year using the curriculum, Mathematics Vision Project, which was the curriculum she used with her focal class for the study, an accelerated eighth grade class. Having led the piloting of MVP, Selena knew the curriculum well.

Selena's focal class had 26 students. The class typically met for 50 minutes. On the day of video recording, the period lasted 42 minutes due to a shortened schedule. Students in Selena's class reflected the demographics of the population of Sierra Middle School: 56% Hispanic or Latino, 37% Asian, over 10% of whom spoke Vietnamese at home, 2% Filipino, 2% two or more races, and 1% white. Eighty-three percent were officially designated "socioeconomically disadvantaged." Twenty-one percent were classified English learners.

The lesson Selena planned, enacted, and reflected on for the study was a review lesson on graphing linear equations. Though the students were in a unit on transformational geometry, Selena wanted to see what their retention on system of equations would be. In the reflection interview Selena explained, "How much are they retaining of what they were taught a month ago? When they take their SBAC Assessment that's what's going to be checked; it's a comprehensive summative assessment." Selena's students were going to take the SBAC math assessment in two weeks' time from the

lesson that was video recorded for the study.

Contextual information about the lesson was garnered from the researcher's observation of the lesson and interview conversation with Selena, not the lesson planning template, since Selena provided no response to the lesson planning template. The lesson was built around a word problem, "The Two Storage Tanks Task." In the problem, one storage tank loses water at a steady, continuous (or uniform) rate while the other storage tank gains water at a steady, continuous (uniform) rate. Before any gaining or losing begins, the tanks start with different amounts of water in them. The rates at which the gaining or the losing occurs differ. Students are expected to determine when the two tanks will contain the same amount of water.

The lesson was intended for students to engage mathematical practice standard 2: Reason abstractly and quantitatively. During reflection Selena explained that the abstract part of the lesson was that "rate of change" being "uniform" means that slope is uniform, which means it has be a straight line, a linear equation. The quantitative part of the lesson was using a lattice point(s) and figuring out the slope of each line.

Posing practice. Overall, evidence of Selena's planning for, enacting, and reflecting on posing located her practice on the continuum at the relational (4) level, flexible posing. At the relational (4) level of practice in posing, respondents demonstrate skillful, strategic flexibility in their questioning. They demonstrate skill in matching their questioning to specific purpose, which they can explain. The purposes of their questioning include eliciting a range of responses, particularly misconceptions about the learning target. They ask how and why questions and questions designed to foreground

students' metacognition. At the relational level of practice, their posing reflects questions from a strategic mix of levels within a taxonomy such as Bloom's, Costa's or Webb's.

Planning. Selena's evidence of planning—for posing, pausing, and probing—was the most incomplete of all subjects who engaged with the P-P-P Assessment. The items design of the P-P-P Assessment sought evidence of planning from: (a) the first interview with study subjects, the intake/planning interview, or SSIP-1, (b) the pre-lesson enactment survey, (c) the lesson planning template, and (d) the final "reflection session" interview, or SSIP-2. Selena did not provide a response to the lesson planning template. Responses to this item were intended to serve as the most direct observation of the teachers' lesson planning performance. Survey and interview responses about lesson planning are self-reported.

Selena also did not provide a response to the pre-lesson enactment survey. Therefore, no self-reported data were collected about the frequency of Selena's plans for posing, pausing, and probing (never, rarely, often, always) via the survey's items 12, 27, 40 that targeted this information about planning. Not responding to that survey also means that Selena did not provide a response to item 41, which was intended to discover if teachers planned specific pausing routines for their students. Item 41 was another Likert-scaled item: "When students need 'wait-time', I create specific routines for them [never, rarely, often, always]."

Evidence of Selena's planning for posing, pausing, and probing came only from her interview responses to the intake/planning session interview and the reflection session interview, which included the VSR protocol.

The self-report evidence of Selena's planning for posing located her practice of this facet of posing at the multistructural (3) level. At this level, respondents plan questions they consider checks for understanding of the lesson's objective. Selena's self-report suggested that she had more in mind than only checking for understanding as she planned, but her description of what she actually did not was not a good enough match with the description of relational (4) planning for posing to locate her planning for posing there.

During the intake/planning session interview Selena said:

> My "first priority" [in planning for that prep] is where their understanding is, checking for understanding. My job in posing questions is to help students see connections. They should see that in what I said yesterday and what I'm saying today there is a connection. It is my duty to bring that out [in my lessons]."

In order to do this for the lesson that was video recorded Selena reported during the reflection session interview:

> I did the problem in my mind before I started the lesson and that was my lesson plan. I looked for all the things I wanted the students to be looking for: graph, line, *x*-axis. I quickly listed all that I was looking for. My questioning has do with how to train the students to look at the graph in a certain way.

Though this latter evidence suggests that Selena plans questions as more than just checks for understanding, there is no "direct" observation of her doing so, as such a response to the lesson planning template might provide.

Enactment. Analysis of video enactment evidence—the recording and the transcript of the recording—supported locating Selena's enactment of posing at the relational (4) level, flexible posing. Consonant with practice at that level, Selena (a) matched questions

and questioning "delivery" to suit a variety of purposes tied to the learning target, (b) posed a mix of questions—including higher-level (according to Webb's DOK or taxonomies such as Bloom's or Costa's), open-ended, and how or why questions; and (c) enacted a lesson in which the activities and pacing clearly reflected that teacher decisions were contingent upon student responses to questions posed.

Reflection. Analysis of the transcript of Selena's response to the VSR protocol determined that her reflecting on posing locates her practice of this facet of posing at the relational (4) level. Selena was able to reflect on perceived effects of changing questions and/or questioning strategies. For example, while unpacking an episode of instruction during VSR Selena both critiqued her initial pose and explained why she handled his response the way she did:

> I asked, "What is the equation?" I should have asked, "What is the *slope-intercept form of a linear* equation?" So when he started to say *f* of *n*, which is basically for a function, I stopped him. I could have allowed him to make that mistake and then gone through that, but I had to literally stop him because I didn't want him to make more mistakes and go with that flow. Next time I would make sure to start off by asking, "What is the slope-intercept form…"

Note that in her reflection Selena takes responsibility for having posed an imprecise question not well-aligned to her purposes for posing it. This is reflective of relational (4) level reflection on posing.

Selena's reflection on posing supported that she was aware of the purposes behind her questioning and that she matched the kinds of questions she asked to these specific purposes, a distinguishing characteristic of relational (4) level posing. During VSR, Selena explained her tactical decision to change from using open-ended questions to using closed-ended questions at a critical juncture in the lesson:

> For the purposes of this lesson I needed them to funnel down to a certain way of seeing. The open-ended question did not bring out what it was supposed to bring out. I wanted them to see the difference, "Hey this scale is different. That scale is different. That's all. The focus is only on that and nothing more. My questioning style at that point was a way of saying, "We're going to do it this way, okay."

Finally, analysis of the transcript of her response to the VSR protocol revealed that Selena was able to suggest several "next steps" likely to support better posing with specificity, indicating relational (4) level reflection on posing.

Pausing practice. Analysis of the body of evidence on Selena's planning for, enacting, and reflecting on pausing located her practice at the multistructural (3) level, intentional, supported pausing. As previously mentioned, at this level of pausing practice, the notion of differentiated pausing closely tied to purpose and context is not of priming concern to respondents. They can, and do, identify and explain reasons for pausing. At the multistructural (3) level of practice, during lessons teachers set up and carry out procedures to protect wait time. Supported pauses longer than three seconds occur in the classrooms of teachers who are at the multistructural (3) level of pausing practice.

Planning. No evidence of planning for pausing was elicited by Selena's engagement with the P-P-P Assessment. Selena did not complete the lesson planning template or the pre-lesson enactment survey, both of which sought to elicit evidence regarding planning for pausing. Without a response, no inferences can be drawn to locate her practice in planning for pausing.

Enactment. Live observation by the researcher, video recording of the lesson, transcripts of the video recorded lesson, and Selena's response to the VSR protocol were

used to locate Selena's pausing enactment. Analysis of the body of enactment evidence supported locating Selena's practice of this facet of pausing at the multistructural (3) level, intentional, supported pausing. At this level, respondents tend to enact routines that especially protect wait time when one student is "in the spotlight" or "on the hook" in whole class or small group settings. Respondents tend to enact "pair shares" and "table talk" as "go to" pausing moves. Directing students to "talk at your tables" as a form of non-silent pausing is a "go-to move" of Selena's. Selena asserted this during VSR and it was corroborated by video evidence. Selena directed students to talk at their tables three times during the lesson.

Video enactment evidence showed that Selena demonstrated verbal and nonverbal support of pausing during the lesson, such as when Selena protected Gunther's think time by saying, "Let Gunther answer." Video enactment evidence revealed Selena carefully and intentionally gave the most and longest pauses to students who needed them most, such as when Ivana was "in the spotlight." Selena carefully facilitated and protected pausing for Ivana's—and the class's—benefit. During a series of public poses and probes to Ivana's thinking during whole class instruction, Selena uncovered a mistake Ivana had made and orchestrated exploration of the mistake during at a pivotal point in the lesson.

A noteworthy pattern in Selena's pausing enactment is that she gave and supported longer pauses to individual students who had been bounced to and that were in the spotlight than the pauses she orchestrated for the whole class immediately after posing a question and before calling on anyone. In this regard, Selena's pausing enactment

demonstrated an unevenness that is not reflective of adjusting for students' needs. Students need more of a pause after a whole class pose.

Reflection. Analysis of Selena's reflection on pausing elicited by the VSR protocol supported locating her skill in this facet of practice at the unistructural (2) level. Respondents at the unistructural level are able to reflect on specific opportunities for pausing that they missed and offer next steps to try to improve their pausing. They are able to reflect on why pausing is important for students. Also, at this level, respondents may want to avoid stretches of silence. These respondents may feel fear, distrust, discomfort, or pain about silence in class, especially in response to questions they have posed. During VSR, Selena identified specific opportunities where she could have supported pausing more. This example is from the transcript of Selena's unpacking of a video clip of whole class instruction:

> So here is an opportunity for them to talk and the talking didn't happen for too long. Therefore pausing wasn't too effective. I have to be quiet like mouse and allow them to think.

During VSR, Selena reflected on why pausing is essential for lessons, important for students, and valuable to teachers:

> In this particular class I didn't give too much time for pausing, but given a chance pausing is a very important factor because you have to allow the students to think, think on their own or even talk amongst each other. That is also pausing time for a teacher. You have to give that pausing moment for any good lesson to happen.

Selena expressed being uncomfortable with whole class silences. "Pausing is difficult—I mean uncomfortable for me as well as for the student, let me be very frank." Selena preferred, she said, "pausing time that means I pose the question, then 'Okay, work in

your group.'" During VSR Selena expressed needing to work on giving pausing time to the whole class immediately after posing a question. My conclusions from analyzing the enactment evidence agree with Selena's assessment of her pausing practice in this regard.

Probing practice. Overall, evidence generated by Selena's engagement with the P-P-P Assessment indicated that Selena's planning for, enacting, and reflecting on probing located her practice at the multistructural (3) level, targeted probing. As previously mentioned, respondents exhibiting targeted probing distinguish between probing to assist teacher decision making and probing to benefit student(s) being probes or any witness learner(s) to the teacher-student(s) probing interaction. Responses to items/tasks indicate probing is potentially valuable to teachers or student decision making. Probes target uncovering misconceptions. The teacher leverages probes and what probes make visible.

Planning. No evidence of Selena's planning for probing was elicited by her engagement with the P-P-P Assessment. Selena did not complete the lesson planning template or the pre-lesson enactment survey, both of which sought to elicit evidence regarding planning for probing. Without a response, no inferences on amounts of the construct an individual possesses can be drawn.

Enactment. Video-based lesson enactment evidence suggests that Selena's enactment of probing matches the description of multistructural (3) level probing on the probing construct map. In line with multistructural (3) probing enactment, records of practice show that Selena probed student responses and thinking in every component of the lesson. Both are indicative of multistructural (3) probing enactment. She also encouraged student-to-student probing. Once in the lesson Selena did this by encouraging a student,

Ivana, to ask her classmates, "Why am I wrong?" Analysis of enactment evidence also demonstrates that what gets elicited via probing gets used by the teacher in attempts to advance the class's learning, a characteristic of multistructural (3) probing.

Reflection. Evidence of reflection on probing elicited by Selena's interaction with the VSR protocol demonstrated that Selena used other FA moves to support probing. This is indicative of multistructural (3) probing. Specifically, during VSR, Selena reported she "gave a little priming, so that later on probing can be helpful." Analysis of lesson enactment evidence bears this out. A distinguishing characteristic of multistructural (3) probing is that respondents are able to distinguish between probing to assist teacher decision making and probing to benefit student(s) being probed or any witness learner(s) witnessing the teacher-student(s) probing interaction. During VSR Selena referred to "incremental probing" that's not just for the recipient of the probing. "I need to see," Selena said, "whether that statement made any dent on anyone."

Selena's reflection during VSR revealed she had a clear sense of what the probing needed to do in the lesson to help students understand "The Two Storage Tanks Task." With great specificity Selena outlined the connections students needed to make regarding rate of change, slope, linear equations, graphs, and lattice points. She asserted, "teachers should probe and ask questions in increments. And when you are asking incremental questions, the questions have to be connected." This reflection suggests that Selena may have expectations of student progression of understanding for the concepts in play in the Two Tanks Task. But without evidence of lesson planning ahead of the lesson there is no way to know if Selena would—as respondents who are multistructural-level in their

probing practices do— "plan specific probes for different points in the lesson and the probes reveal teacher's expectations of progression." Overall, analysis of the transcripts of Selena's response to the VSR protocol indicate locating her reflecting on probing at the multistructural (3) level.

Areas for growth. A priority area for growth concerns the pausing that happens immediately after posing a question to the whole class, before any student is called on, while all students are still "on the hook," so to speak. Because the FA moves inter-relate, Selena's attention to pausing—and particularly how better pausing practices serve to enhance the collection of soft data garnered from other FA moves used in combination with pausing—will likely help Selena better leverage FA moves to advance student learning and optimize her instructional decision making.

Posing. Analysis of the evidence regarding Selena's posing so that it is flexible, functions well in eliciting a range of responses, and is purposeful, shows that Selena knows where she is going with her posing and adjusts when the students' responses are not quite what she expected. An area for growth is for her to consider how other FA moves, particularly pausing, can be aligned to support students' needs. For example, where can Selena be orchestrating different kinds of pausing—besides her "go to" pausing moves of "pair and share" and "talk in your groups" —to support students with questions she knows are more cognitively demanding than others she will ask in the lesson? How does her posing expect progression in students' thinking and how can the other FA moves be synergistically used to support all students' deeper engagement with the content and concepts of the learning target?

Pausing. As identified in the opening to this section, a priority area for growth concerns the pausing that happens immediately after posing a question to the whole class, before any student is called on. Considering how students might be recruited and supported to play roles in efforts to focus on improving the enactment of this kind of pausing is recommended.

Probing. An area for growth is encouraging and scaffolding student-to-student probing. The '"Ask, 'Why I am wrong?' Episode" with Ivana was evidence of actions Selena takes to acculturate students to making mistakes and to get them to probe their own work processes and thinking. It is critically important to continue this kind of probing-related practice. It is recommended that Selena reflect on how she could leverage what she is already doing to get students to probe their own work to support her students in developing the practice of student-to-student probing.

Planning. Lesson planning that commits to differentiated pausing tied to where and how Selena anticipates students will have trouble with the material is recommended. Ideas for differentiating pausing could involve students more actively so that students are not just "recipients" of the FA moves, the ones whom the moves are enacted "upon." Involving students in supporting, even initiating and orchestrating, FA moves will support the enactment of developmentally appropriate pedagogical practices for middle school students. In the middle grades, students are ready to assert increasing autonomy and assume greater responsibility for their learning. Teachers' efforts to align their pedagogical practices with research-supported developmental changes that occur in adolescent cognition and social behavior (Krajcik & Czerniak, 2007; Eccles, Midgley, &

Adler, 1984; Eccles, Lord, & Midgley, 1991) are likely to increase middle grade learners' engagement in and learning from lessons. In Selena's case this could mean focusing on planning ways to scaffold and foster students' roles in initiating, carrying out, and analyzing the class's routines and procedures around pausing.

Enacting. An area for growth for Selena to consider is to intentionally use FA moves synergistically to support student engagement in scaffolded metacognition. Selena's deep content knowledge, long experience teaching in middle grade students of this population, substantial pedagogical content knowledge, and demonstrated skill in improvising make this recommendation a sensible "next step" for focusing her efforts to improve lesson enactment.

Reflecting. Continued reflection on moments where students' responses did not meet teacher expectations for their responses—and what this might suggest about ways to incorporate FA moves into instruction, including encouragement and scaffolding of students to enact FA moves—is recommended. Reflecting on practice with others has a role to play. Since learning is a social process, and because Selena is such an experienced and skillful teacher, and since she demonstrates comfort, competency, authenticity and ease in talking about her practice, Selena's engaging in collaborative reflection with colleagues will serve to both deepen Selena's insights and spark quality learning in her colleagues.

Discussion and Conclusion

The process of locating the teachers' practices of formative assessment in the facets and along the dimensions hypothesized in this study revealed insight into the constructs

targeted by the P-P-P Assessment: teacher posing, pausing, and probing. Locating the levels of Leila's, Lavinia's Jessica's, Aaron's, Eliza's, and Selena's posing, pausing, and probing on the construct maps based on evidence elicited by the P-P-P Assessment also raises questions about the P-P-P Assessment's items design and outcome space and their functioning. I address these and other related aspects next.

Extent of alignment with the construct maps. For this group of six teachers, evidence of practice elicited by the P-P-P Assessment that was relevant to the constructs it targeted mapped almost exclusively to only two of five levels articulated by its construct maps and outcome space: the unistructural (2) and multistructural (3) levels of posing, pausing, and probing. Given the subjects' years of teaching experience, this comports with my expectations for teachers' performances on the P-P-P Assessment. Pre-structural (1) level practice of formative assessment, for example, is more likely to be observed in pre-service teacher populations.

In this study, the only locations of pre-structural level of practice determined were those of a teacher in her second year of teaching. Evidence of Jessica's *reflection on posing* and *planning for probing* elicited through her engagement with the P-P-P Assessment was interpreted to indicate pre-structural level practice for those two facets along those two particular dimensions. No teacher's practice for a dimension "overall," i.e., evaluations that encompassed the evidence for the planning, enacting, and reflecting facets all together, was determined to be located at the pre-structural (1) level.

In the sample obtained for this study, no evidence of extended abstract (level 5) practice was elicited by the P-P-P Assessment either. This aligned with my expectations,

since practice of teacher posing, pausing and probing at the extended abstract level would require planning, enacting, and reflecting on the moves and productively integrating them with known and empirically validated student learning progressions. None of the teachers in the study were using curriculum that explicitly supported them in aligning pedagogical practices—let alone the FA moves as a particular family of pedagogical practices—with empirically validated student learning progressions.

In only one instance was evidence of a dimension of practice elicited by the P-P-P Assessment found to map to the relational (4) level. The dimension was posing. Selena, the teacher with the most years teaching mathematics at the middle school level and with, arguably, the deepest pedagogical content knowledge too, was the only teacher whose engagement with the P-P-P Assessment exhibited solid practice of a dimension of formative assessment as hypothesized in this study at the relational (4) level.

Lavinia's performances related to posing came close, and contained evidence of some instances of enactment above the multistructural (3) level of posing, but did not reach the relational (4) level of posing practice as defined in the current version of the posing construct map. Given the central role of questioning in the practice of formative assessment (Heritage & Heritage, 2013), it is not surprising that the highest levels of teacher practice evidenced were in the posing dimension and were exhibited by teachers with the greatest years of experience teaching children amongst the teachers in the study sample.

Challenges. In mapping evidence of teacher practice to the construct maps I faced two kinds of challenges: (a) challenges related primarily to methods of evidence

collection and interpretation, and (b) challenges related to definition of the constructs themselves. The first kind of challenge arose when considering evidence of lesson enactment. Two issues surfaced: (a) "weighing" teacher performance during small group interactions and whole class instruction and (b) recognizing that the items design and the outcome space for the P-P-P Assessment might not consistently work to elicit the critical information that could reveal the significance of teacher re-visits to an individual student or to a group of students.

Teacher-student interactions in small group and whole class configurations. Using the current outcome space design for the P-P-P Assessment, interpretation of enactment evidence demonstrated the pattern that teacher performance of posing, pausing, and probing moves during interactions with students during small group configuration were more advanced than the corresponding enactments during whole class configuration. This makes intuitive sense and might be explained by cognitive load theory. In contexts of whole class configuration teachers may be attempting to juggle more inputs than they deal with during small group configuration.

However, this noticeable pattern of teacher performance also raises questions about how to fairly collect and interpret evidence of teacher-student interactions vis-a-vis the moves in both configurations when both are recognized as important to instruction and student learning, such questions as:

- Should the items design "balance" the collection of evidence from both configurations when the amount time a teacher engaged in each configuration is far from "balanced"?

- Should the outcome space take into account consideration of such an imbalance?
- Should a teacher who probed well during the only—and very brief—small group interaction she engaged in have that "weigh" equally with her (less sophisticated) probing during whole class configuration?
- What should be done in cases where the difference in teacher performance of the moves according to the scoring guides in the two configurations is widely discrepant?

These and other related questions will be important to consider in future iterations of the P-P-P Assessment.

Surfacing evidence of teacher re-visits and their potential significance. The second challenge of this type, a challenge related primarily to methods of evidence collection and interpretation, that arose while I was locating evidence of teachers' practices of the moves on the construct maps concerned teacher re-visits to students. The issue of a teacher re-visiting a student or group of students concerns all three moves targeted by this study. The "teacher re-visits" issue is particularly salient during lesson enactment, although it could be present in the planning and reflecting facets of every dimension too.

I illustrate the concern with an example regarding teacher probing during small group configuration. A single video clip of a teacher probing an individual student during small group work may not convey that this was the teacher's fourth visit to this group and that the probing the teacher is doing during this particular visit represents a progression in her probing over the visits. The VSR protocol may not reliably function to elicit the

sophistication of teacher practice represented in that single clip. If a teacher is much stronger in the enacting facet, for example, than he is in the reflecting facet, he may leave out contextual information that reveals the significance of his probing in this fourth visit. What if the VSR protocol does not even elicit, in this hypothetical but entirely realistic example, the important fact that this was the teacher's fourth visit to that group? These are the sorts of concerns related to "teacher re-visits" that challenge attempts to assess teacher posing, pausing, and probing reliably and validly.

Need to incorporate craft aspects of practice and reconsider the reflection facet. The second kind of challenge I faced while locating evidence of teacher practice of posing, pausing, and probing on the construct maps relates to how the constructs themselves were defined in this study. The process raised two types of these concerns: (a) concerns about how "craft aspects" of practice relate to definitions of the constructs and (b) concerns oriented toward how the facets of planning, enactment, and reflection—particularly reflection—contribute to defining the constructs.

By "craft aspects" I mean how teachers manipulate their body positions, body language, facial expressions, rate of speech, volume of speech, eye contact with students, and attempts at humor. These do matter for connecting with students and enacting the moves productively. Yet they are not incorporated into the construct maps except for an indirect reference in the description of probing at the multistructural (3) level of practice: "[Respondents] tend to enact probing that targets either movement toward the learning goal or to influence student(s)' affective state(s)."

Future iterations of the P-P-P Assessment should particularly take into account

Lavinia's and Eliza's reflections on posing and probing. Both raised issues of the significance of craft aspects of the FA moves. This empirical evidence should inform the honing of the definitions of these constructs.

The process of locating the empirical evidence elicited by the teachers' engagement with the P-P-P Assessment on the construct maps raised questions about how best to express the phases of planning, enacting, and reflecting in the definitions of the constructs themselves. The descriptions of the reflection facet of each dimension of FA practice were especially challenged by attempts to accommodate the empirical evidence elicited by the P-P-P Assessment. I was challenged to locate on the construct maps teachers who may not have been as "next steps"-oriented as the definition of the constructs presumed (and the items design of the P-P-P Assessment sought evidence for through the VSR protocol), and yet who unpacked their practices during VSR with detailed sophistication. This arose most strongly when considering and attempting to locate Lavinia's reflection.

Finally, Jessica's case raised a concern about the three facets in terms of their role in defining the constructs, a concern that may or may not be mirrored or amplified as attempts to locate greater numbers of teachers' practices of formative assessment on the construct maps are conducted in the future. Evidence of Jessica's practice in the dimension of pausing was determined to be at a different level in each of the three facets: her planning for pausing was located at the unistructural (2) level, her enacting pausing was located at the multistructural (3) level, and her reflection on pausing was located at the pre-structural (1) level. What might these three different locations suggest, if

anything, about the roles the facets play in defining a dimension of formative assessment practice?

Jessica was the only teacher of the six whose locations for a single dimension varied like this on the relevant construct map for the facets. Until greater numbers of teachers are assessed with the P-P-P Assessment it is not possible to make sound sense of what this may or may not imply, if anything.

Opportunities. The process of locating evidence of teachers' practices on the posing, pausing, and probing construct maps enabled me to confirm several hypotheses about the moves. The most significant of these are as follows: first, that teachers do rely on using "go to" moves; second, the process confirmed that teacher action to foster greater student involvement and autonomy regarding the FA moves is indicative of higher-level practice of formative assessment as conceptualized in this study; and third, that the purposes teachers express for their planning, enacting, and reflecting on the moves would give meaningful insight into where evidence of their practice is located on the construct maps. As more teachers in different contexts engage with the P-P-P Assessment other noteworthy findings are likely to occur.

Chapter 6: Significance, Limitations, Future Directions, and Implications

This chapter speaks to the significance of the study, its limitations, possible future directions, and implications of the work. First, I highlight the significance of the P-P-P Assessment's alignment with the principles of educational assessment advocated by an expert committee of the National Research Council (Pellegrino et al., 2001). Second, I foreground noteworthy aspects of the individual formative feedback that resulted from the subjects' engagement with the P-P-P Assessment. Third, I address limitations of the study.

Then I discuss future directions, including what might result from connecting this study and future work related to teacher learning progression in formative assessment with research and professional development projects such as ones that have been led by Rich Lehrer and Leona Schauble. The projects I refer to are ones that have examined how teacher practices change as teachers become familiar with specific student learning progressions and learn to leverage their knowledge of specific learning progressions instructionally. Finally, I discuss implications of the present study to professional development efforts that involve preservice, induction, and inservice teacher populations and the educational leaders and instructional coaches who work with them.

Design of P-P-P Assessment Aligns with NRC's Assessment Triangle

The approach this study took lends significance to the work. By employing the first three building blocks—*construct maps*, *items design*, and *outcome space*—of Wilson's Constructing Measures (CM) framework to inform both the methodology of the study and the design of the P-P-P Assessment, this performance-based assessment of teachers'

posing, pausing, and probing aligns with the elements and principles of educational assessment recommended by a committee of experts of the National Research Council. The elements and principles of educational assessment I am referring to are expressed schematically through a model, known as the Assessment Triangle, and its vertices of *cognition*, *observation*, and *interpretation* (Pellegrino et al., 2001). (See Figure 1 on page 16.)

To foreground the significance of this alignment, it is necessary to outline the elements and principles the NRC recommends for educational assessment (and represents through the Assessment Triangle), connect them to the CM framework, and identify how they were instantiated in this study. I begin first, however, by addressing the strengths of following the NRC's recommendations for educational assessment.

Significance of the alignment, strengths of following NRC's recommendations. The significance of having aligned the design process of the P-P-P Assessment with the concepts represented by the Assessment Triangle in this study derives from the fact that when the NRC's principles are followed, and the elements—as represented by the cognition, observation, and interpretation vertices—soundly interrelate with one another and work in synchrony, the inferences that can be drawn from the assessment are more likely to be meaningful. The better the three elements work in synchrony, the more likely the inferences will be meaningful, and, generally speaking, likelihood that other important concerns related to the assessment are improved will be much greater. These concerns include issues of quality, reliability, and validity.

This is because the concepts *meaningfulness*, *quality*, *reliability* and *validity* in

relation to an assessment are interdependent. An assessment designer should embark on the design and development of an assessment with acute awareness of this inter-dependency. An assessment designer should also work wisely to integrate the components of the assessment that are the designer's attempts to give form and substance to the concepts represented by the *cognition*, *observation* and *interpretation* vertices of the Assessment Triangle. The goal in seeking the integrated functioning of these elements is to strengthen the evidence-based arguments upon which the inferential links of this inter-dependency rest. In the evidence-based tradition of assessment, the quality of the evidence used in these evidence-based arguments is a fundamental and continuous concern. The design of the P-P-P Assessment belongs to this tradition.

How alignment was achieved. The sequential, iterative, and connected nature of the design process for the P-P-P Assessment contributed to its alignment with principles of educational assessment recommended by the NRC. This section summarizes the design process's linkages to the NRC's Assessment Triangle.

First, as recommended, and corresponding to the first vertex of the Assessment Triangle, the theories of *cognition* about the target of inference—teachers' posing, pausing, and probing—were research-based, up-to-date, grounded-in-empirical evidence, and reflective of a developmental stance on learning. These theories of *cognition* were advanced through the creation, utilization, and revision of three *construct maps* that were essential to the design, pilot, and analysis of the P-P-P Assessment.

Second, systematic methods for *observation*, the second vertex of the Assessment Triangle, were determined and expressed via the *items design* of the P-P-P Assessment.

Most significantly, the *items design* was anchored in three performance tasks for the teachers who took the assessment. There was one performance task for each phase of the cycle of inquiry—planning, enacting, and reflecting—that teachers, as professionals, engage in as they practice. The three facets of the practice of formative assessment hypothesized in this study—the facets of planning, enacting, and reflecting—are linked to these phases. This is one foundational way the items design is intentionally related to the construct maps, which express the theories of cognition upon which the assessment is based. In the conceptual language of the Assessment Triangle, this was one method by which I linked the *observation* vertex of P-P-P Assessment to the *cognition* vertex.

Third, methods for the *interpretation* of observations were decided through the creation of the P-P-P Assessment's *outcome space* and articulated through three general scoring guides. This resulted in one scoring guide for each of the constructs targeted by the P-P-P Assessment: one scoring guide for posing, one for pausing, and one for probing. The scoring guides were explicitly linked to the construct maps, which I revisited and revised in light of empirical evidence elicited by the teachers' participation in the study. (The timeline of and details about the revisions to the construct maps were a topic of chapter two.) In the conceptual terms of the NRC Assessment Triangle, this was one method by which I worked to link the *interpretation* vertex to the *observation* vertex in light of the *cognition* vertex. My revisions to the construct maps reflected assertions of the NRC, that "to have an effective assessment...it will almost certainly be necessary for [assessment] developers to go around the assessment triangle several times, looking for mismatches and refining the elements to achieve consistency" (Pellegrino et al., 2001, p.

51).

Purposes. Quality assessments are designed to serve a purpose within a particular context or specified range of contexts (Pellegrino et al., 2001; Wilson, 2005). One aim of working to achieve synchrony amongst the inter-relation of the elements of the Assessment Triangle—and amongst the building blocks of the CM framework—is to better serve the purposes of the assessment within the contexts for which the assessment is being developed. Though this study was carried out in the context of middle school mathematics classrooms, a longer range goal of the endeavor to which this study belongs is to explore ways in which the formative assessment of teachers' practices of FA in several subject area disciplines outside of mathematics could be meaningfully conceptualized and conducted. The actions I took to intentionally and sensibly interrelate the construct maps, items design, and outcome space—which correspond to the cognition, observation, and interpretation vertices of the Assessment Triangle—were undertaken to serve the purpose of the P-P-P Assessment.

The purpose of designing the P-P-P Assessment for my dissertation study was to make ground-breaking progress toward a longer range goal of being able to validly and reliably evaluate three facets of teacher practice of formative assessment—planning for, enactment of, and reflecting on—practice of FA along several hypothesized dimensions: posing, pausing, and probing among them. Concomitant with this purpose was to carry out the design of the P-P-P Assessment and the work of the study in such ways that evidence-based individualized formative feedback for teachers could be generated. I address the significance of the individualized formative feedback that resulted from my

study next.

Outcome of Teacher Engagement with the P-P-P Assessment: Individual Formative Feedback

The generation of individual formative feedback to the teachers who engaged the P-P-P Assessment, which was presented in chapter five, is noteworthy for several reasons. This section outlines the reasons.

First, the feedback itself serves as a proof of concept that it could be done. Second, the existence of the individual formative feedback also demonstrates that the P-P-P Assessment, within the context of this study, achieved one of its purposes, which was to play a pivotal role in the generation of targeted feedback for each study participant. The feedback intended to fall within each respondent's zone of proximal development in the practice of formative assessment as conceptualized in this study.

Analysis to determine if this intention was realized, however, was not conducted. Exploration of the usefulness of the feedback to the subjects in the study—or any other qualities or characteristics of the feedback presented in chapter five—was not conducted either. While important, and possible to conduct in the future with new study subjects, examinations of this type were not within the scope of the present study.

Third, the significance of the individual formative feedback presented in chapter five relates to the framing of this dissertation study as making a contribution toward the articulation of a *teacher* learning progression (TLP) in the domain of formative assessment. Arguments in the field for developing and supporting student learning progressions (SLPs), arguments that outline the significance of student learning progressions to student learning, apply as well to the idea of teacher learning progressions

and teacher learning. These arguments highlight the significance of the individual formative feedback presented in chapter five.

When referring to *student* learning progressions (SLPs), Heritage (2008), Alonzo (2011), Black, Wilson, and Yao (2011) and others (See, e.g., Shavelson et al., 2010; and Furtak et al., 2008) have argued that learning progressions are central to the practice of formative assessment. In particular, Heritage's reasons (2008) for calling for the development of student learning progressions apply to the need for and significance of developing *teacher* learning progressions—especially in the domain of formative assessment.

Heritage (2008) has argued that learning progressions are "foundational" to three "key elements" of the practice of formative assessment as it benefits learners (p. 5). One of these "key elements" is that learning progressions are "foundational" to "provid[ing] feedback to students" (p. 5). The other two "key elements" of formative assessment for which the existence and use of learning progressions are "foundational" are: (a) "elicit[ing] evidence about learning to close the gap between current and desired performance" and (b) "involv[ing] students in the assessment and learning process" (Heritage, 2008, p. 5).

Within the framing of this study as contributing toward the articulation of a *teacher* learning progression in the domain of teacher practice of formative assessment, the individual formative feedback presented in chapter five, therefore, stands as an example of how a nascent and emerging teacher learner progression can be used to provide feedback to teachers. Although a teacher learning progression in formative assessment

has not yet been fully articulated, nor validated—this study contributes *toward* a TLP—the feedback of chapter five is noteworthy because it exists in relation to an emerging teacher learning progression.

I expect that as teacher learning progression in the practice of formative assessment is more fully explicated, based on future studies and more empirical evidence, and as validation work proceeds, both the process of generating individual formative feedback for teachers and the content of that feedback will evolve. The feedback presented in this study could serve a role in future efforts to improve how formative feedback to teachers in this domain is generated, articulated, and presented.

In the future, the feedback presented in this study could also be compared with feedback generated by teachers' engagement with future assessments developed using more refined articulations of teacher learning progression in the domain of formative assessment, assessments that might or might not be related to the P-P-P Assessment. Such comparisons could be practically useful to stakeholders such as instructional coaches and preservice teacher educators and important to research-based efforts to improve the quality of feedback available to teachers.

Limitations

This section outlines the limitations of the present study. Limitations derive from four areas: (a) sample size and selection bias, (b) lack of anchoring learning progressions, (c) use of video, and (d) curriculum and content effects, which can be considered major moderating variables.

Sample size and selection bias. Although significant and useful findings result from

studies with an *n* of six—the number of subjects in this study—chances are that a sample of this size, when combined with selection bias, will not represent a normal distribution. For this study, selection bias contributed to little evidence of subjects' practices being found to match the lowest level of practice described on each of the construct maps and scoring guides. Therefore, how teachers' engagement with the P-P-P Assessment functioned to help generate feedback to teachers whose practices are found to be located at the lowest level in the dimensions of posing, pausing, and probing was untested, a limitation on the generalizability of the findings of this study.

Instructional coaches and preservice educators who work with teachers whose practices of posing, pausing, and probing match the lowest level articulated on the construct maps and scoring guides may find the feedback in chapter five has limited relevance to their contexts. On the other hand, coaches and preservice educators who work with teachers whose practices of posing, pausing, and probing match the middle levels articulated on the construct maps and scoring guides—the unistructural (2) and multistructural (3) levels—are likely to find the feedback presented in chapter five relevant to their work.

The same holds true regarding the higher levels of practice articulated in the present study. Little evidence of subjects' practices was found to match the higher levels of practice described on each of the construct maps and scoring guides, i.e., only one teacher exhibited relational (4) level practice in one facet of one dimension: enactment of posing. No evidence at all was found in the study for the highest level, extended abstract (5). Therefore, the way in which teachers' engagement with the P-P-P Assessment functioned

to help generate feedback to teachers whose practices are found to be located at the higher levels in the dimensions of posing, pausing, and probing remains untested, a limitation on the generalizability of the findings of this study.

Lack of anchoring learning progressions. The lack of empirically validated student learning progressions for the teachers to use as they planned, enacted, and reflected is a limitation of the present study. Because the study subjects did not have access to relevant student learning progressions, it is impossible to determine whether or not evidence of their practice would have been found to match higher levels of practice on the construct maps and scoring guides if a relevant student learning progression were available to them. I surmise that some of the study subjects would have been able to leverage their knowledge of said learning progression to good effect while other study subjects would not have been able to do. The present study, however, can make no claims related to teacher practice and student learning progressions, since no SLPs, anchoring or otherwise, were employed in the study.

Use of video. Video-based limitations include that video recording cannot provide an objective view of what occurs in classrooms (Fadde & Zhou, 2014; van Es et al., 2015). While studies support the use of video in evaluation of teachers' practice (Greenberg, Kane, & Thal, 2015), including teachers having reported that administrator evaluations were more supportive when conducted by video rather than through live observation, there are negatives to including video in processes of teacher evaluation of practice. One such negative is Sherin and Han's (2004) finding that a focus on evaluation when teachers are reviewing their video often prompts teachers to make quick judgments about

what is viewed on video, without careful consideration of what was taking place and why. A focus on "fixing practice" can overtake conversation about "learning about one's practice" (Sherin & Dyer, 2017).

It is not possible to know to what extent this stance was in play for the teachers in the present study. Conducting think alouds and taking more actions to collect validity evidence based on response processes could have shed light on the question: did teachers' tendencies to focus on "fixing practice" overtake their careful consideration and exploration of what was taking place and why as they reviewed video clips of their own instruction during the video-stimulated recall protocol? However, no think alouds were conducted, and few items of the P-P-P Assessment targeted collecting validity evidence based on responses processes (although the 10-item Exit Survey did). This is a limitation of the present study.

Curriculum and content effects. Other limitations of the study derive from curriculum and content effects. The study could not control for the curriculum used by teachers. Though all the teachers taught in the same school district, not all the curriculum was the same. Two of the teachers planned, enacted, and reflected lessons using one curriculum. The other four teachers used another curriculum. Some evidence suggested that more proficient teacher probing was associated with use of the curriculum that was better at providing example questions and probes to the teachers.

Teacher practice of planning, enacting, and reflecting on posing, pausing, and probing could have been substantially influenced by content effects too. This is an area ripe for future exploration.

Future Directions

Future work should continue to establish the validity and reliability arguments for evaluating teachers' formative assessment practices by means of the P-P-P Assessment. Such work should be conducted in order to support the wider use and adoption of the P-P-P Assessment as a method for (a) improving the quality of formative feedback available to teachers in the critical domain of formative assessment and (b) continuing the development of teacher learning progression in this domain. Another important direction for future work related to this study is to combine this work, with its focus on teacher development of formative assessment practices, with research on teachers' use of student learning progressions.

Validity studies. Validity studies that delve further into evidence based on response processes should be conducted. This work would entail the use of cognitive labs and verbal reports (Ericsson & Simon, 1993). Additionally, collecting validity evidence based on the examination of the internal structure of the P-P-P Assessment should be pursued. This involves creating and analyzing Wright maps and conducting item analysis, including, as needed, analyses of differential item functioning. Such work, conceptually, would be employing the fourth block of Wilson's building blocks: measurement model.

Additionally, to further support arguments for validity, investigations into scores and score interpretations from the P-P-P Assessment with relations to external variables should be done. Schoenfeld's TRU Math scheme and the FARROP are recommended as potential observation tools to employ when relating external variables to the P-P-P Assessment.

Reliability studies. Though RQ3 asked, "Can teacher practices of posing, pausing, and probing along any or all of these proposed dimensions of practice be reliably and validly evaluated?" the present study did not elicit quantitative evidence concerning reliability related to the P-P-P Assessment. As such, future studies involving the P-P-P Assessment should incorporate investigations of reliability and should focus on the inter-rater reliability of the functioning of the scoring guides and scoring protocols on video-based evidence.

Combine research on TLPs with research on SLPs. For several reasons, future work should combine research on teacher learning progression in the domain of formative assessment with research that features an anchoring student learning progression, or several validated student learning progressions. Without the presence of a validated student learning progression in a research project that features the P-P-P Assessment (or its construct maps), the descriptions of the highest levels of practice hypothesized in the construct maps for posing, pausing, and probing cannot be tested, nor revised, based upon empirical evidence. This is because all the descriptions of extended abstract (5) level practice of the dimensions of FA practice hypothesized in this study require teachers to productively leverage a known student learning progression during planning, enactment, or reflection. Therefore, without there being a known student learning progression in play, evidence of teacher practice cannot, by definition, be interpreted to indicate performance at the extended abstract (5) level.

Potential to accelerate teacher development. A second reason to combine research on TLPs with SLPs is that knowledge of both are needed to in order to best support—

perhaps even accelerate—teacher learning and student learning. I will illustrate the need for this combination by highlighting Kim's (2010) findings about teachers' improving skill in learning to interpret student responses to items during a research project that had a goal of establishing a learning progression.

Kim found that participating teachers began by treating student responses to the assessment items as right or wrong (Kim, 2010). Over time, however, teachers learned to interpret student responses to items within the learning progression being established. Kim (2010) found that as teachers grew more attuned to the nuances of student responses, they began to (1) assess responses based on the milestones of the construct map, (2) invite explicit comparisons among forms of reasoning described by the construct, and (3) press for understanding on how higher level forms of reasoning increased the scope and/or precision of an explanation.

This "press for understanding" is related to probing, though not necessarily a proxy for probing. As a researcher of teacher learning progression in the domain of formative assessment, my reading of Kim's findings is that as teachers learned more about the learning progression being established, they became better at practicing formative assessment. This is as I would surmise, and a reason that future directions related to this work should include the combination of studying TLPs and SLPs together.

Role of video in future work. Moreover, the work that will study TLP with SLP should thoughtfully incorporate video of teacher practice that features student expression. I recommend this for two reasons. One reason is based on Kim's work; the other is based on Sherin and Dyer's experiences with teachers in video "clubs" (Sherin & Dyer, 2017).

Kim and the research team acknowledged a particular use of video in their study in contributing to the changes observed in teacher practices over time. Specifically, the research team assisted teachers in creating video-annotated construct maps of student understanding. The researchers video recorded "formative assessment conversations that exemplified how teachers might employ discussion of the items to support students' conceptual change" (Lehrer, 2012, p. 181). These video-clip annotated construct maps "clarified how student talk and activity correspond[ed] to particular levels of one or more constructs" (Lehrer, 2012, p. 181-2). Significantly, the "annotations helped teachers view learning performances more dynamically as the formative assessment conversation unfolded in the classrooms" (Lehrer, 2012, pp. 181-2).

Without specifically referencing established learning progressions, qualitative researchers Sherin and Dyer (2107) describe a value of teacher participation in informal video "clubs". They describe video club participation as an activity that helps teachers to pay closer, more nuanced attention to student thinking and communication, and that the group discussion of teacher practice encourages teachers to innovate instructional practices that begin to carry over to their subsequent teaching (Sherin & van Es, 2009).

In the following excerpt, as Sherin and Dyer (2017) explain their methodology during video club meetings, note how they seem to be helping teachers (a) to explore the meaningfulness behind perceived similarities and differences of student thinking and (b) to delve into where student thinking might be in relation to an undefined progression, rather than categorize student ideas, thinking, and responses dichotomously:

> In video clubs, we encourage teachers to look beyond whether a student's idea is correct or incorrect and to try to understand how the student might have

> developed that idea, and how different students' ideas are related. "Where do you think Zach may have gotten the idea that the slope was zero?" "Do you think Hannah and Mateo are making the same point?" These questions often lead teachers to develop new instructional practices based on the explanations they discuss. (p. 53)

How much better might the video club discussion become if there were a common framework, such as the FA moves framework, to discuss teacher practice and if, for the subject matter content in question, there were an established (student) learning progression? This question points to why I recommend that future research combine TLP, SLP, and careful use of video clips of teachers' own practices.

Implications

This study has implications for the professional development of teachers. The empirical evidence from this study could be utilized in professional applications that serve preservice, induction, and inservice teacher populations. Stakeholders expected to find the work of this study relevant to their contexts include preservice teacher educators, teacher induction specialists, instructional coaches, principals, educational professionals concerned about teacher evaluation, educational researchers, and those who provide professional development to inservice teachers.

Much of the value of the study resides in the approach taken in designing the P-P-P Assessment, an approach specifically taken to support the generation of quality formative feedback to teachers interested in developing their skills in the domain of formative assessment. Though the sample pool comprised only middle school mathematics teachers, implications of the work extend beyond the context of middle school mathematics instruction. The data and findings of the study should also serve to support

the resourcing of further efforts to explore the multi-dimensionality of teacher practices of formative assessment from a moves-based perspective.

The moves-based conceptualization of formative assessment employed in the study shows promise in helping teachers realize more equitable instruction during class. The implications of this should not be understated. Lenses and frameworks used to examine teaching are not neutral. The moves-based framing of formative assessment has been shown in this study and in previous work to foreground the use of academic language in lessons and to play an important role in supporting teachers in making more productive improvisational moves during class time. When teachers of different disciplines can use the work of this study to self-assess their practice, maintain meaningful inter-disciplinary conversations, and use the work and their conversations to create more equitable opportunities for student learning across departments, schools, and districts, a powerful possible application of this work will have come to fruition.

References

Abedi, J. (2010). Research and recommendations for formative assessment with ELLs. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 181–197). New York: Routledge.

Allal, L., & Ducrey, G. P. (2000). Assessment of—or in—the zone of proximal development. *Learning and Instruction*, *10*(2), 137-152.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421.

American Educational Research Association (AERA, APA, NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, L.W. & Krathwohl, D. R. (Eds.), Airasian, P.W., Cruikshank, K. A., Mayer, R. E., Pintrich, P.R., Raths, J., & Wittrock, M. C. (2001). *A Taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete edition). New York: Longman.

Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of Formative Assessment*. New York: Routledge.

Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic English: Implications for K-12 English language learners.* Arlington, VA: The George Washington University Center for Equity and Excellence in Education. Retrieved from https://pdfs.semanticscholar.org/76b8/476fd601e434e53b6c6edd2855b5e2fe1b45.pdf

Arter, J. A. (April, 1998). *Teaching about performance assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Athanases, S. Z. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *Elementary School Journal*, 94, 421-439.

Bailey, A.L., & Heritage, M. (2008) *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum.* Thousand Oaks, CA: Sage/Corwin Press.

Balfanz, R. (2009). "Putting Middle Grades Students on the Graduation Path." Policy and Practice Brief. Westerville, OH: National Middle School Association. Retrieved from http://www.amle.org/portals/0/pdf/articles/Policy_Brief_Balfanz.pdf

Ball, D. L., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *The Elementary School Journal*, *109*(5), 458-474.

Black, P., Wilson, M., & Yao, S.Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspective, 9*(2-3), 71-123.

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), 139–144.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives handbook I: The cognitive domain.* New York: David McKay.

Boaler, J. (2002). Learning from teaching: Exploring the relationship between reform curriculum and equity. *Journal for Research in Mathematics Education*, *33*(4), 239-258.

Boaler, J., & Sengupta-Irving, T. (2016). The many colors of algebra: The impact of equity focused teaching upon student learning and engagement. *Journal of Mathematical Behavior*, 41, 179-190. doi:10.1016/j.jmathb.2015.10.007

Bodrova, E. & Leong, D.J. (2007). *Tools of the mind*. Columbus, OH: Prentice Hall.

Boerst, T. A., Sleep, L., Ball, D. L., & Bass, H. (2011). Preparing teachers to lead mathematics discussions. *Teachers College Record*, *113*(12), 2844-2877.

Borko, H. & Putnam, R. (1996). Learning to teach. In R. Calfee and D. Berliner (Eds.), *Handbook of Educational Psychology* (pp. 673-708). New York: MacMillan.

Borko, H., Roberts, S. A., & Shavelson, R. (2008). Teachers' decision making: From Alan J. Bishop to today. In P. Clarkson & N. Presmeg (Eds.), *Critical issues in mathematics education: Major contributions of Alan Bishop* (pp. 37-67). New York: Springer.

Boston, M. D., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of Instructional Quality Assessment Toolkit* (CSE Tech. Rep. No. 672). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator*, *3*(2), 154-175.

Brookhart, S. M., & Nitko, A. J. (2015). *Educational assessment of students,* 7th edition. New York: Pearson.

Bruner, J. S. (1960/2009). *The process of education*. Cambridge, MA: Harvard University Press.

Calderhead, J. (1981). Stimulated recall: A method for research on teaching. *British Journal of Educational Psychology*, *51*, 211-217.

California Department of Education (CDE). (2010 & 2013). California Common Core State Standards Mathematics, *Electronic Edition.* Retrieved from http://www.cde.ca.gov/be/st/ss/documents/ccssmathstandardaug2013.PDF

Chung, R. R. (2008). Beyond assessment: Performance assessments in teacher education. *Teacher Education Quarterly*, *35*(1), 7-28.

Common Core State Standards Initiative (CCSSI). (2010). *Common Core State Standards for Mathematics.* Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers. http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. (CPRE Report). Philadelphia, PA: Consortium for Policy Research in Education.

Council of Chief State School Officers. (2008). Formative assessment: Examples of practice. A work product initiated and led by Caroline Wylie, ETS, for the Formative Assessment for Students and Teachers (FAST) Collaborative. Washington, DC: Author.

Council of Chief State School Officers. (2013, April). Interstate Teacher Assessment and Support Consortium InTASC *Model Core Teaching Standards and Learning Progressions for Teachers 1.0: A Resource for Ongoing Teacher Development*. Washington, DC: Author.

Darling-Hammond, L. (2013). *Getting teacher evaluation right*. New York: Teachers College Press.

Darling-Hammond, L. & Adamson, F. (Eds.) (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. San Francisco: Jossey-Bass.

Darling-Hammond, L. Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, *25*(3), 179-204. doi:10.1007/s11092-013-9163-0

Darling-Hammond, L. & Wei, R. C. (2009). Teacher preparation and teacher learning: A changing policy landscape. In G. Sykes (Ed.), *The handbook of education policy research* (pp. 613-636). Washington DC: American Education Research Association.

Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction. (CPRE Report). Philadelphia, PA: Consortium for Policy Research in Education.

Darragh, L. (2013). Constructing confidence and identities of belonging in mathematics at the transition to secondary school. *Research in Mathematics Education*, *15*(3), 215-229.

Duckor, B. (2006). Measuring measuring: An item response theory approach (doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No 3253841.

Duckor, B. (2005, May). Thinking about the act of measuring: The development of a theory of the construct. Individual poster presented at the 2nd Annual Meeting of the Center for Assessment and Evaluation of Student Learning Conference, Santa Rosa, California. Available from Center for Assessment and Evaluation of Student Learning at http://www.caesl.org/conference2005/brent_sm.pdf

Duckor, B. & Holmberg, C. (in press). Focusing on moves-based formative assessment to increase equity of voice in middle school mathematics: A case for Video-based professional development. In S. B. Martens & M. M. Caskey (Series Ed.), *The Handbook of Research in Middle Level Education: Preparing middle level educators for 21st century schools: Enduring beliefs, changing times, evolving practices.* Washington, DC: AERA Press.

Duckor, B. & Holmberg, C. (2017). *Mastering formative assessment moves: 7 High-leverage practices to advance student learning*. Alexandria, VA: ASCD.

Duckor, B., Holmberg, C., & Rossi Becker, J. (2017). Making moves: Formative assessment in mathematics. *Mathematics Teaching in the Middle School*, *22*(6), 334-342.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606-609.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123-182.

Eccles, J. S., Lord, S., & Midgley, C. (1991). What are we doing to early adolescents? The impact of educational contexts on early adolescents. *American Journal of Education*, *99*(4), 521–542.

Eccles, J., Midgley, C., & Adler, T. F. (1984). Grade-related changes in the school environment: Effects on achievement motivation. In J.G. Nicholls (Ed.), *The development of achievement motivation* (pp. 283–331). Greenwich, CT: JAI Press.

Ericsson, K. A., & Simon, H. A. (1999). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Massachusetts Institute of Technology Press.

Fadde, P. J. & Zhou, T. (2014). Technical considerations and issues in recording and producing classroom video. In B. Calandra & P. Rich (Eds.), *Digital video for teacher education: Research and practice* (201-216). New York: Routledge.

Ferguson, R. F., and Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.) *Designing teacher evaluation systems*[1]*: New guidance from the measures of effective teaching project*. Hoboken, NJ: John Wiley and Sons, 98-133.

Finkelstein, N., Fong, A., Tiffany-Morales, J., Shields, P. & Hoang, M. (2012). *College bound in middle school and high school? How math course sequences matter*. Retrieved from the Center for the Future of Teaching and Learning at WestEd website: https://www.wested.org/wp-content/files_mf/1399319766319 21CFTL_MathPatterns_Main_Report.pdf

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Research-based recommendations for instruction and academic interventions: Practical guidelines for the education of English Language Learners.* (Vol. 1). Portsmouth, NH: Center on Instruction.

Furtak, E. M., Ruiz-Primo, M.A., Shemwell, J. Ayala, C., Brandon, P., Shavelson, R., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning." *Applied Measurement in Education*, *21*(4), 360–89.

Gotwals, A. W., & Alonzo, A. C. (2012). Introduction. In *Learning progressions in science* (pp. 3-12). Sense Publishers.

Graesser, A.C., & Person, N.K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104-137.

Greenberg, M. Kane, T. & Thal, D. (2015, April). When teachers choose: Fairness and authenticity in teacher-initiated classroom observations. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Guerra, P. L. & Nelson, S. W. (2009). Changing professional practice requires changing beliefs. *Phi Delta Kappan*, *90*(5), 354-359.

Gunckel, K. L., Mohan, L., Covitt, B. A., & Anderson, C. W. (2012). Addressing challenges in developing learning progressions for environmental science literacy. In H. Andrade, & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 39-75). Rotterdam, Netherlands: SensePublishers.

Hakuta, K. (2013). *Assessment of content and language in light of the new standards: Challenges and opportunities for English learners*. Princeton, NJ: The Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/rsc/pdf/hakuta_assessment_content_language_standards_challenges_opportunities.pdf

Hakuta, K. (2014). Assessment of content and language in light of the new standards: Challenges and opportunities for English language learners. *The Journal of Negro Education, 83*(4), 433-441. doi:10.7709/jnegroeducation.83.4.0433

Hakuta, K. (2017). Retrieved from http://web.stanford.edu/~hakuta/

Hambleton, R. K. & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, *5*, 1-16.

Henning, J. E., McKeny, T., Foley, G. D., & Balong, M. (2012). Mathematics discussions by design: creating opportunities for purposeful participation. *Journal of Mathematics Teacher Education*, *15*(6), 453-479.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do?. *Phi Delta Kappan*, *89*(2), 140.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Washington, DC: The Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2008/Learning_Progressions_Supporting_2008.pdf

Heritage, M. H. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin.

Heritage, M., & Heritage, J. (2013). Teacher questioning: The epicenter of instruction and assessment. *Applied Measurement in Education*, *26*(3), 176-190.

Hill, H.C., Blunk, M.L., Charalambous, C. Y., Lewis, J.M., Phelps, G.C., Sleep, L. & Ball, D.L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430-511, DOI: 10.1080/07370000802177235

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation.

Kennedy, C., Brown, N., Draney, K., & Wilson, M. (2005). Using progress variables and embedded assessments to improve teaching and learning. Annual Meeting of the American Education Research Association, Montreal, Canada.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568-589.

Kim, M.-J. (2010). A case of collaboration between researchers and teachers mediated by boundary objects. In R. Lehrer & M. Wilson (Chairs), *Assessing a multidimensional learning progression: Psychometric modeling and brokering professional development.* Symposium conducted at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, M.-J. & Lehrer, R. (2015). Using learning progressions to design instructional trajectories. In C. Suurtamm (Ed.), *Annual Perspectives in Mathematics Education (APME) 2015*:*Assessment to Enhance Teaching and Learning*. (pp. 27-38). Reston, VA: National Council of Teachers of Mathematics.

Klein, R. (2015, May 5). "In 10 years America's classrooms are going to be much more diverse than they are now." *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/2015/05/07/classroom-demographics-2025_n_7175760.html

Krajcik, J. & Czerniak, C. (2007). *Teaching science in elementary and middle school: A project-based approach.* New York: Routledge.

LaParo, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409-426.

Lehrer, R. (2012). A learning progression emerges in a trading zone of professional community and identity. Retrieved from http://www.uwyo.edu/wisdome/_files/documents/lehrer.pdf

Lehrer, R., Kim, M.-J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning*, *12*(3), 195-216.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education*, *96*(4), 701-724.

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. P. Maloney, H. Confrey, & K. H. Nguyen (Eds.) *Learning over time: Learning trajectories in mathematics education* (pp. 31-59). Charlotte, NC: Information Age Publishing.

Lehrer, R. & Kim, M. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116–133.

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195–216.

Linn, R. L. (1989). *Educational measurement*. New York: ACE.

Lustick, D., & Sykes, G. (2006). National Board Certification as professional development: What are teachers learning? Education Policy Analysis Archives, 14(5). Retrieved from http://epaa.asu.edu/ojs/article/view/76

Marton, F. (1981). Phenomenography—describing conceptions of the world around us. *Instructional science*, *10*(2), 177-200.

Masters, G., & Forster, M. (1997). *Mapping literacy achievement: Results of the 1996 National School English Literacy Survey*. Canberra, Australia: Department of employment, Education, Training and Youth Affairs.

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment*, *13*, 267–300.

Mehan, H. (1979). *Learning lessons*. Cambridge, MA: Harvard University Press.

Mehan, H. (1979). 'What time is it, Denise?": Asking known information questions in classroom discourse. *Theory into Practice*, *18*(4), 285-294.

Milanowski, A. (April, 2011). Validity research on teacher evaluation systems based on the framework for teaching. Individual paper presented at the American Educational Research Association conference, New Orleans, Louisiana.

Molnar, M. (2017, May 24). "Market is booming for digital formative assessments." *EdWeek*. Retrieved from http://www.edweek.org/ew/articles/2017/05/24/market-is-booming-for-digital-formative-assessments.html

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: W. H. Freeman and Company.

Moschkovich, J. N. (2015). Academic literacy in mathematics for English learners. *The Journal of Mathematical Behavior*, *40*, 43-62.

Moschkovich, J. N. (2015). Scaffolding student participation in mathematical practices. *ZDM*, *47*(7), 1067-1078.

Moschkovich, J. N. (2013). Issues regarding the concept of mathematical practices. In *Proficiency and beliefs in learning and teaching mathematics* (pp. 257-275). Rotterdam, Netherlands: Sense Publishers.

Moschkovich, J. (2013). Principles and guidelines for equitable mathematics teaching practices and materials for English language learners. *Journal of Urban Mathematics Education*, *6*(1), 45-57.

Moyer, P. S., & Milewicz, E. (2002). Learning to question: Categories of questioning used by preservice teachers during diagnostic mathematics interviews. *Journal of Mathematics Teacher Education*, *5*(4), 293-315.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

National Committee on Science Education Standards and Assessment, National Research Council. (1996). *National science education standards*. Retrieved from National Academy of Sciences website: http://www.nap.edu/catalog/4962.html

O'Hara, S., Zwiers, J., & Pritchard, R. (2012). Framing the teaching of academic language. Retrieved from https://cset.stanford.edu/sites/default/files/ALLIES%20Brief%20v8.pdf

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education 57*(1), 22-36.

Pellegrino, J.W., Chudowsky, N., Glaser, R. & National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Popham, W. J. (2008). Transformative assessment. Alexandria, VA: Association for Supervision and Curriculum Development

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4-13.

Rasch, G. (1960). *Probabalistic models for some intelligence and attainment tests*. [Reprinted by University of Chicago Press, 1980].

Rudner, L. M. & Boston, C. (1994). Performance assessment. *The ERIC Review*, *3*(1), 2-12.

Ruiz-Primo, M. A., Solano-Flores, G., & Li, M. (2014). Formative assessment as a process of interaction through language. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing Assessment for Quality Learning* (pp. 265-282). Netherlands: Springer.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

Sahin, A., & Kulm, G. (2008). Sixth grade mathematics teachers' intentions and use of probing, guiding, and factual questions. *Journal of Mathematics Teacher Education*, *11*(3), 221-241.

Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal*, *45*(3), 669-700.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 5.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM*, *45*(4), 607-621.

Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productively intertwined. *Educational Researcher*, *43*(8), 404-412.

Schoenfeld, A. H. (2015). Summative and formative assessments in mathematics supporting the goals of the common core standards. *Theory Into Practice*, *54*(3), 183-194.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., ... & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632-654.

Sherin, M. G., & Dyer, E. B. (2017). Teacher self-captured video: Learning to see. *Phi Delta Kappan*, *98*(7), 49-54.

Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, *20*(2), 163-183.

Sherin, M.G. & van Es., E.A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20-37.

Smarter Balanced Assessment Consortium (SBAC). (2017). History. Retrieved from Smarter Balanced Assessment Consortium website: http://www.smarterbalanced.org/about/history/

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, *4*(1-2), 1-98.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, *46*(6), 610-631.

Spanos, G., Rhodes, N. C., & Dale, T. C., & Crandall, J.(1988). Linguistic features of mathematical problem solving: Insights and applications. *Linguistic and Cultural Influences on Learning Mathematics*, 221-240.

van Es, E.A., Stockero, S., Sherin, M.G., van Zoest, L., & Dyer, E.A. (2015). Making the most of teacher self-captured video. *Mathematics Teacher Educator, 4* (1), 6-19.

Wiggins, G. & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F.K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics* (pp. 1053–98). Charlotte, NC: Information Age Publishing.

Wiliam, D. & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2005). *Constructing measures*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716-730.

Wilson, P. H., Mojica, G. F., & Confrey, J. (2013). Learning trajectories in teacher education: Supporting teachers' understandings of students' mathematical thinking. *The Journal of Mathematical Behavior*, *32*(2), 103-121.

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

Wright, B. D., & Masters, G. N. (1981). *Rating scale analysis*. Chicago: MESA Press.

Wylie, C., Lyon, C., & Formative Assessment for Students and Teachers (FAST) State Collaborative on Assessment and Student Standards. (SCASS). (2013). *Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice*. Washington, DC: Council of Chief State School Officers.

Zwiers, J., O'Hara, S., & Pritchard, R. (2014). *Common Core Standards in diverse classrooms: Essential practices for developing academic language and disciplinary literacy*. Portland, ME: Stenhouse Publishers.

Zwiers, J., O'Hara, S., & Pritchard, R. (2014). Conversing to fortify literacy, language, and learning. *Voices from the Middle*, *22*(1), 10.

Appendices

Appendix A: Construct Maps

Construct Map: Teacher Posing of Questions

High	*Respondents who* integrate relevant features of the context for learning with multiple important purposes for questions (e.g., promoting meta-cognition). They pose questions that size up the context for learning in ways that reflect knowledge of students' development, interests, needs re: learning target(s), and present understandings. They pose questions that relate to the lesson and the unit plan and larger essential questions/big ideas of the discipline. <u>They plan questions</u> that reveal explicit anticipation of where students may/are likely to get stuck or have misconceptions. They plan questions that serve to provide evidence in helping teachers decide which of a few to several specific (and expressed) decisions they might make that are contingent upon students' responses to these questions (e.g., they plan "hinge" questions and post-hinge question pathways for instruction). They plan in ways that encourage student questions to be springboards for discussion. They plan questions that reflect a balance between content-centered instruction with student-centered instruction. <u>They tend to enact lessons</u> that display several ways student responses can be used to further students' own and other students' learning regarding the lesson target. They tend to enact lessons that feature questions that reflect a sensible balance in addressing a variety of learners' needs. <u>They are able to reflect</u> on how questions posed functioned to elicit evidence of student understanding in relation to lesson objectives/target(s)	Integrative posing (Extended Abstract) 5	*Responses to items/tasks* indicate flexibility in posing to adjust to students' learning edges in real-time in relation to learning goals. Questions posed leverage a range of student responses (including student questions) in ways that elicit evidence of having furthered students' present understandings in relation to the lesson target and/or essential question/big idea of the discipline. Responses to items/tasks show that respondent has anticipated student pit stops and bottlenecks typical of learning progression of concept/skill/understanding. Observation of teaching shows student responses being used in a variety of ways, including changing the direction of the lesson and/or pausing an activity.

	of instruction.		
	Respondents who demonstrate flexibility in their questioning. They demonstrate an awareness of the variety of purposes of their questions and the need to match kinds of questions to specific purposes. They plan a variety of questions designed to elicit a wide range of responses, including misconceptions and "unorthodox" responses. They plan carefully sequenced repetition of key questions. They plan supports/scaffolds for questions. Then tend to enact lessons in which activities and pacing clearly reflect teacher decisions that are contingent upon student responses to questions posed about the learning target. They are able to reflect on perceived effects of changing questions and/or questioning strategies. They are able to suggest several "next steps" likely to support improved posing. They do so from many perspectives and with specificity.	Flexible posing (Relational) 4	*Responses to items/tasks* indicate posing of *how* and *why* questions and questions from a mix of Webb's DOK or other taxonomic levels (e.g. Bloom's, Costa's). Observation of teaching will likely show changing questioning strategies in response to student(s) response(s). Observation of teaching may show students playing significant roles in posing questions. Observation of teaching will show many questions that serve to highlight connecting students' prior knowledge and experiences with present efforts to engage with and "reach" the learning target.
	*Respondents who*se purposes for questioning seem to be to get students to say what the respondent-as-teacher is thinking (rather than eliciting from students a range of responses, including unknown responses, responses surprising to the teacher). They plan questions they consider checks for understanding of the lesson's objective. Then tend to enact lessons with high percentages of close-ended questions. They tend to enact lessons in which scenarios arise where students are expected to guess what the teacher is thinking, even when doing so appears more a hindrance to than a help	Constrained posing (Multistructural) 3	*Responses to item/tasks* indicate posing a high percentage of, or posing only, *what/when/where*, fact recall, and lower-level questions (on Webb's DOK, Bloom's taxonomies etc.). Questions planned connect to learning target. Observation of teaching shows questions posed as checks for understanding procedures and concepts tied to the learning target. Observation of teaching shows questions that seek to elicit students' prior knowledge. Observation of teaching reveals questions posed seldom elicit a wide range of responses.

Low	regarding students' advancement toward the learning target. They are able to reflect on several aims of improving posing. Their reflection includes specific suggestions for alternate poses to try.		
	Respondents who demonstrate through their questioning a primary focus on orchestrating student behavior, not necessarily learning (activity-based posing). They may not be able to make student thinking visible through questions they pose. They plan questions that do not reveal clear priorities in the purposes of posing questions. As they plan, they experience challenges in deciding what content is most important to ask about and when. They tend to enact teacher-centered lessons that do not reflect an underlying pedagogical structure dependent upon student responses to curricular content. They are able to reflect on benefits that might accrue from using a questioning scheme.	Posing to manage (Unistructural) 2	*Responses to items/tasks* indicate posing to manage/control students, e.g., "Do you have a pencil? Are your books open to page 39?" Planned questions do not express recognizable coherence or organizing principle. Observation of teaching and questions reveal imbalance of focus between activity/behavior and learning target.
	Respondents who give directions to students and whose actions can be interpreted as attempting to pour content into students' minds without eliciting from the students where their current understandings are. They may plan questions not well-crafted to elicit evidence of student understanding in relation to instructional goal(s). They tend to enact lessons that do not invite or incorporate students' prior knowledge. They are able to "reflect" through descriptions of their instruction that do not push to analysis.	Pre-posing (Prestructural) 1	*Responses to items/tasks* indicate no questions aligned with lesson target are posed by the teacher. Planned questions may or may not align with lesson target(s). Observation of teaching may show random or arbitrary questions.

Construct Map: Teacher Pausing

High	*Respondents who* adapt pausing procedures to a variety of cognitive and affective needs that are tied to demands of instruction. They can explain how their pausing moves benefit more systematic and equitable evidence gathering (e.g., pausing's role in increasing sample size), class/instructional participation, and decision making. They plan several different kinds of pausing moves in relation to what is being asked of which students, why, and when in the lesson and learning cycle. They tend to enact pausing moves that offer choices to students and that facilitate independence, inter-dependence, and self-assessment and that include scaffolding as needed. They are able to reflect in detail with sophistication on the potential effects of alternative pausing moves for individual students, groups of students (for a variety of groupings), and the whole class in relation to the learning target and in light of what they know about each individual and each group of learners. They are able to reflect on how students have contributed to pausing moves.	Contextualized & differentiated pausing (Extended Abstract) 5	*Responses to items/tasks* indicate pausing tailored to individual and group needs (e.g., ELs, students with 504 plans) and responsive to changing contexts. Pausing reflects purposeful attention in decision-making to student, context, and curriculum. Observation of teaching reveals contextualized use of "think time" based on explicit curricular challenges and/or student learning styles.
	Respondents who say they need to protect students' think time. They express value in protecting and using pausing to better their instructional decision making. They may prioritize their own processing needs amongst values	Strategic pausing (Relational) 4	*Responses to items/tasks* indicate a variety of approaches to pausing and that pausing approach chosen/taken intentionally relates to student(s) involved or expectations of student performance (e.g., if more

	they express. They plan scaffolding for pausing that fosters student access to materials as needed to support thinking during pauses in relation to the learning goal. They tend to enact pausing moves of different quantities and qualities. They tend to enact a mix of "quiet", "noisy", "active", "still", "individual" and "group" pausing moves and routines throughout a lesson to fit lesson goals, pacing, and learners' needs. They are able to reflect and suggest different ways to increase students' roles in pausing moves. They are able to reflect on how they have benefited from pausing and on how they—and/or their instructional decision making—might benefit further from pausing.		detailed student responses are expected, longer think times may be necessary depending on students' prior knowledge, skill level, and scaffolds available). Observation of teaching reveals strategic use of "think time".
Low	*Respondents who* can define reasons for pausing. The notion of differentiated pausing closely tied to purpose and context is not of primary concern to them. They plan pausing moves to foster student participation and improved equity of participation. They plan pausing moves to elicit better quality responses from students. They tend to enact routines that especially protect wait time when one student is "in the spotlight" or "on the hook" in whole class or small group settings. They tend to enact "pair shares" and "table talk" as "go to" pausing moves. They are able to reflect on motivations for increasing flexibility and strategic use of	Intentional, supported pausing (Multistructural) 3	*Responses to items/tasks* indicate pauses a few seconds (1-3) to several seconds (3-8+) long and procedures for setting up and carrying out pausing. Observation of teaching reveal consistent and careful, intentional use of "wait time".

	pausing moves (and accompanying scaffolds for) and offer suggestions on how.		
	Respondents who want to avoid stretches of silence. May feel fear, distrust, discomfort, pain about silence in class, especially in response to questions they have posed/directions they have given. They plan lessons with few or no explicit pausing moves or routines. They tend to enact lessons and student-and-teacher "dialogues"/exchanges that reflect their discomfort with silence, especially with moments of "whole class silence" after they have posed a question. They are able to reflect on specific opportunities for pausing that they missed and offer next steps to try to improve their pausing. They are able to reflect on why pausing is important for students.	Unintentional, unsupported Pausing (Unistructural) 2	*Responses to items* indicate pauses that begin are ended prematurely by teacher action. May interrupt or miss pausing cues. Not well organized or modulated. Confound with discipline or time on task. Observation of teaching shows "walkabout tactics".
	Respondents who are not attentive to how much and why think time matters. They plan lessons without explicit pauses or pausing moves. They tend to conduct classroom procedures (such as call-and-response patterns) without protecting student wait/think time. They are able to reflect on why it might be necessary to change their practices related pausing.	Pre-pausing (Prestructural) 1	*Responses to items/tasks* indicate pausing does not occur. Observation of teaching shows no pausing.

Construct Map: Teacher Probing

High	*Respondents who* can explain what they anticipate to happen as a result of their probing and why they assert this. They communicate concern for responding productively to student responses they cannot anticipate and can name strategies for doing so. They plan probes with relevant knowledge of students in mind. They plan ways to support students probing each other's thinking and performances. They tend to enact lessons that reflect multi-faceted purposes for probing. They tend to enact lessons that integrate other FA moves to support probing. They are able to reflect with sophistication on present probing practices and suggest specific next steps that encourage student independence and interdependence related to probing and are likely to aid student growth toward learning target(s).	Progressive & responsive probing (Extended Abstract) 5	*Responses to item/tasks* indicate pattern(s) to probing that include productive teacher responses to information newly elicited by probing and that is incorporated into further probing. Planned probes reflect use of learning progressions (in relation to learning targets) in their prioritization. Observation of teaching shows productive handling of surprise or "unorthodox" responses. Observation of teaching reveals that what probing elicits is progressively used to advance student responses toward the learning target. Observations of teaching show probing that features detailed and relevant knowledge of students, context, and curriculum. Observations reveal a copasetic and productive balance achieved "reconciling" probing delivery/technique with students' affective states and learning goals.
	Respondents who can describe the value of specific probes/probing moves in relation to their purposes. Respondents who adjust their probing methods or strategies according to incoming evidence ("evidence" that may or may not be gathered/processed systematically or strategically) and in light of the goal(s) regarding the learning target. They plan probes that should serve to elicit an intentional range of responses/performances in order to set up instructional decision making contingent upon what the probes elicit. They plan lessons that incorporate these probes and that can accommodate—or leverage—the implications of what the probes help	Contingent probing (Relational) 4	*Responses to item/tasks* indicate teachers take up evidence of student performance in probing formulation or delivery. A variety of probing moves are demonstrated. Observation of teaching shows that student-to-student (S-2-S) probing occurs and that there are routines, scaffolds, norms around S-2-S probing. Observation of teaching reveals frequent "take up" of students' ideas, exact words, and "presumptions" in the formulation of probes. Observation of teaching may show "extended episodes" of probing that are on topic and on task between teacher and students and students and students in whole class, small group, and one-on-one configurations.

	make visible (e.g., modular, if-then, "flow chart"-like plans for lessons).They plan for student-to-student probing to occur in some components of their lessons. They tend to enact probing that results in responses that get used by students and teacher. They are able to reflect in detail on what they are looking/listening for when they probe. They are able to reflect on the affordances and constraints of the multi-faceted purposes of their probing. They are able to reflect on how probing relates with other FA moves and possible implications of these connections for students' learning in relation to the learning target.		
	Respondents who distinguish between probing to assist teacher decision making and probing to benefit student(s) being probed or any witness learner(s) witnessing the teacher-student(s) probing interaction. They plan specific probes for different points in the lesson and the probes reveal the teacher's expectations of progression. They tend to enact lessons that include probing of "correct answers." They tend to enact lessons that include probing in every configuration in the lesson: one-on-one, small group, and whole class. They tend to enact lessons that encourage student-to-student probing. They tend to enact probing that targets either movement toward the learning goal or to influence student(s)' affective state(s). They are able to reflect in ways that articulate "next steps" that incorporate what probes were intended to or did reveal. That is,	Targeted probing (Multistructural) 3	*Responses to items/tasks* indicate probing is potentially valuable to teacher or student decision making. Probes target uncovering misconceptions. Teacher leverages probes and what probes make visible. Observation of teaching show that what gets elicited via probing gets used by the teacher or students in attempts to advance the class's learning of the learning target (e.g., "Sammy said he predicts *x* because *y*. How does that explanation (y) compare with the reasoning behind your prediction?"). Observation of teaching may reveal that a focus on probing depth sacrifices achieving a wide range of information via probing, and constrains the evidence available to inform the teacher's decision making related to whole class snapshots of understanding related to the learning target (e.g., teacher learns a lot about one or two students' understandings, which may or may not be representative of many others' understandings).

	they are able to conjecture on possible, alternate post-probe pathways for instruction. They can speak to how improved probing might improve "options" for pathways for instruction and how this might benefit certain learners or groups of learners (e.g., "stuck" students, students holding certain misconceptions).		
	Respondents who contend the main purposes of probing are to spur student action and to make learners' thinking more visible, though their actions may imply reasons/purposes for probing beyond those. But these reasons/purposes are not expressed explicitly in planning, enacting or reflecting. <u>They plan</u> specific and/or generic probes (e.g., "Why?" "How do you know?"), and often include their "go to" probes. They may plan probes that reveal anticipation of student confusion or areas of challenge for students. They may plan some specific probes tied to content. <u>They tend to enact</u> most of their probing in one component/portion of the lesson. They tend to enact probing to prod student action. They may enact probes tied to content or the learning target. <u>They are able to reflect</u> on lesson and identify missed opportunities for probing or further probing. They are able to reflect and suggest alternate probes that, while related to the learning target, still may be generic or task-focused only.	Task-focused Probing (Unistructural) 2	*Responses to items/tasks* indicate probing relies on generic probing moves, e.g. "Why?" "Say more…" "What do you mean?" as "go to" probes that are beyond "probing to manage" or "probing to engage". Probes may or may not elicit new information from learners. Observation of teaching show that some probes "work" and make some learners' present thinking visible. Observations of teaching may show there is some evidence that what gets elicited via probing gets used by the teacher or student.
Low	*Respondents who* do not generate probes related to the learning target when planning and enacting. <u>They plan</u> lessons absent of probes related to the learning target.	Pre-probing (Prestructural) 1	*Responses to items/tasks* do not plausibly indicate that probing related to the learning target has occurred. Observation of teaching shows probing not related to the learning target.

	They tend to enact lessons where student discourse is not rich and where "discussions", if they occur, exemplify "coverage and review" not processes supporting "uncovering". They are able to reflect and suggest (at least a few) alternate probes relevant to the learning target (that may be generic and/or task-focused only).		

Appendix B: Scoring Guides

Scoring Guide—Posing

Score	Descriptors
5	**Poses questions that size up the context for learning and reflect knowledge of students (including needs), the learning target, and students' present understandings** Teacher/instruction/posing ● anticipates where students typically get stuck in progressing toward learning target ● reflects clear purposes (e.g. revealing misconception(s), promoting metacognition) tied to learning target and big ideas of the discipline ● incorporates a range of student responses (including student questions) to promote focused disequilibrium and new schema development in individual and groups of students ● embodies a balance between content-centered and student-centered instruction ● tailored to individual and group needs (ELs, students with 504 plans) based on explicit curricular challenges ● plans hinge questions and uses knowledge of learning progressions relevant to the curriculum and learners to plan questions ● reflects on how questions posed functioned to elicit evidence of student understanding in relation to the learning target
4	**Flexibly and strategically matches questions and questioning "delivery" to suit a variety of purposes tied to learning target** ● and does so in ways that improve amount and quality of "evidence" of student "understanding" available ● poses a mix of questions—including higher-level (according to Webb's DOK or taxonomies such as Bloom's or Costa's), open-ended, and *how* and *why* questions ● plans and enacts questions that elicit a wide range of responses, including misconceptions and "unorthodox" responses ● adjusts in response to student responses (including student questions) ● plans carefully sequenced repetition of key questions ● incorporates other FA moves to support posing
3	**Exhibits limited range of purposes of, questions for, and responses from posing** ● poses a high percentage of lower-level (according to Bloom's, Webb's DOK, etc.) and closed-ended questions ● falls into "guess what the teacher is thinking" exchanges

	● elicits students' prior knowledge related to learning target ● seldom elicits a wide range of responses ● plan questions as "checks for understanding" ● reflection explores benefits of improving posing and offers specific suggestions how (e.g., use Habits of Mind, Costa's levels)
2	**Poses to manage or control students** ● is not able to make student thinking visible ● is activity-based: e.g., "Are your books open to page 39?" ● does not plan for posing systematically or craft questions to pose using a method/guidance/organizing principle ● plans do not communicate clear/prioritized purposes for posing ● reflects on how to improve posing
1	**No posing plausibly related to learning target occurs** ● may plan questions not well-crafted to elicit student responses related to learning target ● may enact questions that appear arbitrary or random ● reflection may explore possible questions related to learning target and appropriate for students
0	**No response (irrelevant or off-topic)**

Scoring Guide—Pausing

Score	Descriptors
5	**Contextualized use of "think time" based on curricular challenge and/or student learning style/need** Teacher/instruction/pausing ● tailored to individual and group needs (e.g., ELs, students with 504 plans) ● reflects purposeful attention to student, curriculum, and task/prompt in relation to learning target ● plans for and can explain why several different kinds of pausing moves are used with which students, why, and when in the lesson and learning cycle ● reflects on how pausing practices could better serve individual/group needs
4	**Strategic use of "think time" improves student access to curriculum and teacher decision making** ● includes a mix of "quiet", "noisy", "active", "still" "individual", "group" "directed" and "undirected" pausing routines selected to fit learners' needs regarding advancing toward lesson target ● makes own needs for pausing for "think time" a priority ● encourages students' roles/responsibilities regarding pausing ● can explain several benefits of structured pausing and how pausing can influence decision making ● unpacks practices related to pausing from more than one orientation (e.g., learner-focused orientation, assessment-focused, equity-focused)
3	**Intentional use of "wait time" includes routines for non-silent pausing** ● features "pair-shares" and "table talk" as "go to" pausing moves ● demonstrates verbal and nonverbal support of pausing ● whole class silences last from a few to several seconds long ● may include pausing routines inserted on the fly when "not enough hands up" or "too many blank looks" ● pausing moves may appear "one size fits all" ● plans for pausing moves to increase student participation and elicit "better" student responses, though may not be well-articulated on planning documents ● offers suggestions/reasons for improving pausing
2	**Concerned with getting self and students quiet for lengths of time** ● Unplanned, spontaneous pauses "prematurely" ended by teacher ● features no public expressions during class of valuing pausing or

	wait/think time ● may include intermittent pauses that happen incidentally/accidentally ● plans lessons that do not include explicit pausing procedures ● is undermined by teacher discomfort with silence, classroom management skills, or (lack of) confidence in management skills ● planning documents do not anticipate places in the lessons where pauses are needed to support student learning, promote more equitable participation, and increase quality of responses ● reflection may explore reasons for "rushing," may include suggestions for pausing, often tactics (e.g., "I will count.")
1	**No pausing moves are planned or observed during enactment** Teacher ● identifies and reflects on missed opportunities for pausing ● may acknowledge importance of pausing ● may offer ideas/suggestions for how to support/improve pausing

Scoring Guide—Probing

Score	Descriptors
5	**Anticipates where and how students typically get stuck and leverages student responses to advance multiple students' understanding of target content** Teacher/instruction/probing ● can explain where to probe first, next, last and why ● productively handles "surprise" responses ● incorporates detailed and relevant knowledge of students, context, and curriculum ● integrates other FA moves for synergistic effects ● includes student-to-student probing ● reflection features different kinds of "next steps" recommendations for improving practice (e.g., includes class structure-level and/or lesson component-level suggestions as well as alternative probes to try)
4	**Incorporates students' ideas, words and "presumptions" and the responses elicited by probing get "used" by students and teacher** ● adjusts probes to incoming evidence ● probes a range of student responses/performance "levels" to inform decision making ● can explain what they are listening/looking for during probing ● can speak to how probing interrelates with other FA moves and implications for instruction and assisting students ● reflects productive balance with learning goals and students' affective states ● can suggest specific next steps that would use/incorporate student responses to probes ● reflection includes sophisticated examples of alternate probes, suggestions for improved probing
3	**Targets uncovering misconceptions; may sacrifice "breadth" for "depth"** ● probes "correct answers" ● occurs in every configuration (whole class, small group, one-on-one) in the lesson ● exhibits a range of probing moves that suggest different purposes behind probing moves planned and/or enacted ● planned probes reveal expectations of progression or challenging/confusing concepts ● plans verbatim/scripted probes that are specific or in different components of the lesson ● can suggest alternative specific probes tied to learning target

	● how certain learners ("stuck" students, students with particular misconceptions) could benefit from improving probing/revising instruction
2	**Focuses on spurring student action and making student thinking more visible to teacher** ● applies generic "go to" probes (e.g., "Why?" "Say more…") indiscriminately ● occurs mostly (and markedly so) in one component/portion of the lesson ● may plan probes with no clear organizing principle behind them and/or their placement in the lesson ● reflection includes specific/verbatim possible probes tied to the content/"big idea" connected to learning target
1	**No probing plausibly related to learning target occurs** ● during planning, enactment, or reflection ● reflection identifies missed opportunities for probing ● reflection includes suggestions for and probes relevant to the learning target (that may be of dubious quality, e.g. they may be very few, generic or only task-focused)
0	**No response** (irrelevant or off-topic)