Complex Assessments, Teacher Inferences, and Instructional Decision-Making

by

Theodore Alan Barone

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Education in the Graduate Division of the University of California, Berkeley

Committee in charge
Professor Mark Wilson, Chair
Professor David Pearson
Professor Charles Weisselberg

Fall, 2012
Complex Assessments, Teacher Inferences, and Instructional Decision-Making

© 2012

Theodore Alan Barone
Abstract

Complex Assessments, Teacher Inferences, and Instructional Decision-Making

by

Theodore Alan Barone

Doctor of Education

University of California, Berkeley

Professor Mark Wilson, Chair

The purpose of this study is to understand how teachers make sense of data from a complex set of reading assessments and how inferences they derive from the data affect decisions about instruction. There are two key research questions - 1) How does a teacher’s developing expertise with a complex set of assessments (assessment literacy) affect the quality of the inferences they make about student learning?; and 2) How does the extent of a teacher’s Pedagogical Content Knowledge affect the range and quality of instructional decisions made upon the basis of assessment results?

Six English teachers from a middle school and a high school in Northern California administered a set of reading assessments constructed from various core elements in the reading domain. The data generated from the assessments were organized to facilitate analysis of the progress of teacher thinking about each of the different reading dimensions and with assessment and instruction. Measures of assessment literacy, a series of coded interviews and professional development with the teachers provided data about the development of their Pedagogical Content Knowledge (PCK) of reading. Their ability to make sense of this complex set of student reading data to make instructional decisions was examined over time.

It was found that the skill and sophistication of the teacher’s PCK of reading and interpretation of the assessment evidence increased as they delved deeper into close analysis of both the reading measures and the individual student data. The results have implications for a teacher-based “community of judgment” as a key element in effective school accountability systems.
Dedication

To my beloved wife, Mary, who has been unrelenting in her support of this insanity, and to my beautiful children, Nathan and Devon who inspire me to do everything I can to help public education be the incubator of a healthy future for our planet.

Also, to the hosts of dedicated teachers and administrators and parents and students who are critical partners in the creation of that future. I hope that I can make a small and useful contribution.
# Table of Contents

## Chapter 1: Introduction and Problem of Practice  
  p. 1  
  - Theory of Action  
  - Research questions  
  - Conceptual framework  

## Chapter 2: The Knowledge Base  
  p. 3  
  - School Accountability  
  - Pedagogical Content Knowledge  
  - Assessment Literacy  
  - Expertise and quality  
  - Assessment validity  
  - Reading assessment  

## Chapter 3: Methodology and Detailed methods  
  p. 10  
  - Cross-Case Study/Action Research  
  - Research Design  
  - Case Selection  
  - Data Collection  
  - Data Design  
  - Data Analysis  
  - Research Validity and Reliability  

## Chapter 4: Findings  
  p. 18  
  - Profiles of Cases  
    - Sandy  
    - Greg  
    - Betsy  
    - Albert  
    - Janis  
    - Gene  
  - Qualitative Data Analysis  
    - At the beginning - low Pedagogical Content Knowledge  
    - Struggling to make sense  
    - Improvement in analytical depth and sophistication  
      - Cognitive and Metacognitive Strategies  
      - Vocabulary  
    - Students making meaning from text: How precise does it need to be?  
    - Restructuring conceptual frameworks  
    - Growth in PCK, increased motivation to expand instructional strategies  
    - Synthesis of qualitative analysis  
  - Quantitative Data Analysis  
    - Assessment Literacy Inventory  
    - Rubric scores  

## Chapter 5: Conclusion and Implications  
  p. 50
Appendices

1. Interview Questions
2. Record of Participation
3. Assessment Literacy Inventory (ALI)
4. Metacognitive Awareness of Reading Skills Inventory (MARSI)
5. Coding Configurations
6. Professional Development - Reading Domain
7. Professional Development – Essential Research Findings about Effective Strategies for Teaching Reading

Bibliography

Figures and Tables

- Figure 1: Theoretical Impact of PCK and Assessment Literacy on Instructional Decision-making p. 1
- Figure 2: Quality of inferences about student learning p. 2
- Figure 3: Pedagogical Content Knowledge p. 4
- Figure 4: Order of Operation p. 13
- Table A – Rubric for Pedagogical Content Knowledge and Instructional Decisions p. 14
- Table B – Rubric for Assessment Literacy and Inference Quality p. 14
- Table C – Assessment Literacy Inventory Results p. 47
- Table D – Rubric Scores p. 48
- Table E – ALI and Rubric Correlation p. 49
Acknowledgements

So many people contributed to my education over the past several years. Without their support, I don't believe I could have managed to complete this endeavor. Working full time as the Principal of a major high school; going to classes evenings and weekends as a graduate student; being a husband and the father of two teenagers; and trying to maintain my health and fitness can take a severe toll.

Obviously, I acknowledge my wonderful family to whom I have dedicated this dissertation.

Special notice is given to Professor Mark Wilson of the Graduate School of Education at Cal. Mark was the perfect advisor for me. I didn't want a lot of guidance but when I needed it, he always had just the right (and concise) question to ask or recommendation to make. Professor David Pearson and Professor Elfrieda Hiebert shared their extraordinary knowledge about reading with me and were essential mentors throughout my research. Professor Rick Mintrop, Program Director and founder of the Leadership for Equity in Education Program worked hard to keep me focused throughout my studies and played a key role in my evolution by driving me to a scholarly way of thinking. Chuck Weisselberg, Professor of Law at Boalt, stepped in and agreed to serve on my orals and dissertation committees. I'm glad for his friendship. Cynthia Greenleaf of WestEd provided critical support at an early stage of my research. My colleagues in the LEEP 3 cohort – Jivan Dhaliwal, Linda Wells, Kyla Johnson, Sarah Williams and Lisa Brooks – provided a camaraderie and support that made the whole experience richer and more satisfying. They are a great group of educators. Cynthia Coburn taught the course in qualitative research methods which was invaluable in the end, plus she was an outstanding teacher, one of the best I've ever had. Finally, my administrative team at Albany High School – Tami Benau and Susan Charlip - picked up the slack when I had to attend to one thing or another over the course of my studies. And last, but not least, Marla Stephenson, the Superintendent of the Albany Unified School District made the initial recommendation to apply to the LEEP program and she supported me in multiple ways through to its conclusion.

Thank you one and all. I have reached the end of this particular journey in no small measure because of you.
Chapter 1: Introduction and the Problem of Practice

The purpose of this study is to help understand what teachers do to make sense of data from complex assessments and how inferences derived from the data are typically used to make decisions about instruction. The results will likely have implications for a range of educational contexts, including curriculum design, teacher pre-service training, in-service professional development, data warehousing software, and standards and accountability systems.

The study is situated in a societal context of the public’s desire to educate children to be prepared to understand and resolve some of the most vexing problems ever to face humanity (National Commission on Excellence in Education, 1983; National Governors Association, 2010; Vickers, 1996). Constant international conflict, global climate change, the expanding economic disparities between nations, and exponential growth in world population pose significant challenges to the existence of humanity itself. Their solutions will require a high level of creativity, humanity and critical thought but the rapid development of these problems since the early 1900’s has outstripped the ability of educators and traditionally educated youth to meet their complex demands.

In response to these problems, the curriculum standards movement has begun a shift towards complex, integrated skills and knowledge that require a higher degree of critical thinking and problem-solving skills. The recently adopted Common Core is a set of curriculum standards that embody this new approach.

Theory of Action

The theory of action behind this study is that the quality of the inferences made about student learning is dependent upon teacher Pedagogical Content Knowledge and assessment literacy. Pedagogical Content Knowledge (PCK) includes teacher knowledge of the specific subject matter and the ability to design and carry-out the instructional strategies that are appropriate for the topic and the students being taught while assessment literacy (AL) is the ability to use, evaluate and interpret the assessments used to produce the data. As illustrated in Figure 1, this theory of action would predict that as a teacher’s assessment literacy and/or PCK improves, then the quality of his inferences about student learning and subsequent decisions about instruction would also improve.

![Figure 1: Theoretical Impact of PCK and Assessment Literacy on Instructional Decision-making](image)
Research Questions

The following broad research questions, focused specifically on the reading domain, guided this study:

1) How does a teacher’s developing expertise with a complex set of assessments affect the quality of the inferences they make about student learning?
2) How does the extent of a teacher’s Pedagogical Content Knowledge affect the range and quality of instructional decisions made upon the basis of assessment results?

Conceptual Framework

As described previously, the purpose of this study is to attempt to understand how teachers use data derived from complex assessments to make decisions about future teaching. A framework for understanding this sense-making process is graphically represented in Figure 2.

Two conceptual lenses are provided representing the two major knowledge domains of this research - Pedagogical Content Knowledge (PCK) and assessment literacy. As a teacher examines the results of assessments from a complex set of measures, inferences are derived about student learning after being filtered through his Pedagogical Content Knowledge and level of assessment literacy. The degree to which these two knowledge domains overlap has an impact on the quality of inferences a teacher makes about assessment results (In Figure 2, the thickness of the arrows represents quality of inferences about student learning). This affects their ability to make instructional decisions that make sense in the context of the original learning goals. I would also suggest that the teacher’s skill affects the likelihood of conscious decisions about next teaching steps. A teacher with low PCK and assessment literacy is more likely to be compelled to stick to a schedule (i.e. finish the Civil War by Christmas) than figuring out a way to be sure the student has achieved the learning goal.
Chapter 2: The Knowledge Base

To better understand and establish a context for my problem of practice, I consulted the literatures on school accountability, Pedagogical Content Knowledge and assessment literacy, with a particular focus on the teaching and assessment of reading, an examination of expertise, and finally, modern validity theory. A brief overview of each of these elements is followed by a conceptual framework that provides a lens through which we can consider the problem.

School Accountability

The 25-year movement to establish a uniform external accountability system achieved its apex with the No Child Left Behind legislation of 2001. However, this centrally controlled, high-stakes accountability movement is increasingly blamed for failing to prepare America's youth for the problem-solving skills required for the 21st Century, driving teachers out of the profession, and increasing educational inequities (Au, 2007; Booher-Jennings, 2005; Mintrop & Sunderman, 2009). An alternative to this failing system is to refocus on a rigorous, locally oriented accountability system that is centered in classrooms and schools. The research assembled here indicates that an accountability system situated in the local context can be successful when there is a fairly high level of teacher professional capacity and effective systems for monitoring its implementation (Elmore, 2004; Kelley & Protsik, 1997; Wilson, 2004).

In his meta-synthesis of 49 qualitative studies on the impact of high stakes testing on curriculum, Au (2007) found a strong tendency of schools to align their curriculum to the test, particularly in a way that meets testing norms. He found that pedagogy became more teacher-centric as schools strived to cover the breadth of test-required information and procedures. A number of the studies connected high-stakes testing with increased drop-out rates and lower achievement, particularly for working class students and students of color. Booher-Jennings (2005), in his study of the Texas accountability system, sought to understand the mechanisms that teachers and schools used to meet the demands of the system. He discovered a kind of “educational triage” whereby teachers and schools diverted resources to support those students who were most likely to do well on the tests and to those on the “threshold of passing”, the so-called “bubble kids”.

Control of current central accountability systems is held by a group of what Mintrop and Sunderman (2009) call “centrally positioned elites” - which range from profit-oriented corporate CEOs to civil-rights minded politicians - who are able to take advantage of new information and data warehousing technologies that make incentives and sanctions easier to centralize and control. These data warehouses enable the targeting of a small number of performance indicators and monitoring of a large number of students, each of whom is tagged with a variety of identifiers (e.g. ethnicity, gender, home language, etc.) that allow managers to “surgically order sanctions for underperforming units” (p. 353).

There is a strong tendency for such an accountability system to increase the divide between high and low-capacity schools and districts. High capacity schools already have resources to perform at high levels and can take advantage of the
system’s impetus to further improve their instruction and curriculum. These schools are able to focus on loftier educational goals such as critical thinking and student engagement. On the other hand, low-capacity schools lack the resources to respond so they use short term strategies geared to produce immediate improvement on the test which leads to further “rigidity, fragmentation, and deterioration” of the curriculum and, as a result, “students are excluded from intellectually challenging content and learning” (Mintrop & Sunderman, 2009, p. 358).

A sense of individual responsibility and shared expectations within a school community, an important characteristic of a high capacity school, is key to how schools respond to external accountability regimes, according to Elmore in his study of school reform (2004). He examined ways that schools decide to whom they are accountable, for what, and how. He found schools that demonstrate alignment between their internal expectations and personal responsibility tend to be more immune to external influences, thus better able to weather the vagaries of the external accountability regime. He concluded that a strong set of professional norms within a school community, combined with the knowledge and skill necessary to achieve their collective goals, are essential to success with strong external accountability systems.

A strong internal accountability culture as described by Elmore, is dependent on a coherent localized assessment system. Wilson (2004) provides a framework for considering this coherence, describing it as a “community of judgment” that puts teachers and their judgments at the center of the accountability system. It includes a developmental framework for describing levels of student achievement; assessment tasks that provide opportunities for students to demonstrate what they know; “moderation” systems that evaluate teacher judgments; and quality control methods to ensure valid and reliable interpretation of assessment evidence.

**Pedagogical Content Knowledge**

A teacher’s instructional capacity is dependent upon their Pedagogical Content Knowledge (PCK) - the interplay between their subject-area knowledge, their pedagogical skills, and understanding of the context in which the learning takes place (Baxter & Lederman, 2002). PCK provides the foundation of my knowledge base.

PCK is not simply a mixture of its three different elements. As illustrated in Figure 3, it is a compound construct in that the interplay between the three elements is transformative and creates a unique set of knowledge and skills that together are much more than each analyzed separately.

The altered nature of these complex skills and knowledge makes PCK particularly difficult to measure. Most researchers tend to separate their investigations.
along subject matter or pedagogical knowledge lines because it is easier to narrow the scope of one’s focus and “formulate a question far less complex than the form in which the world presents itself in practice” (Shulman, 1986, p. 6). It’s challenging enough to assess content knowledge, but no test or set of observations will show why a teacher chose one strategy over another. “The PCK construct has fuzzy boundaries, demanding unusual and ephemeral clarity on the part of the researcher to assign knowledge to PCK” (Gess-Newsome, 2002, p. 10).

Assessment Literacy
The next major layer in my knowledge base is assessment literacy which represents a teacher’s understanding of the key dimensions of the assessment process, i.e. the selection or development of methods that are appropriate to the subject area and/or student progress, effective administration, and analysis and interpretation of the results, including the grading or decision-making required as an outcome of the assessment (Mertler, 2009).

Assessment literacy is inextricably linked with Pedagogical Content knowledge in that effective teachers are aware of how their theories of learning interact with the occasionally subtle indicators of student achievement. ”Expert teachers are more skilled at monitoring and interpreting the complexity that characterizes (assessment and) instruction." (McMillan, 2003). It allows them to depart from planned instruction as they reflect on the impact of current and planned activities.

As with so many aspects of teaching, there is an intricate web of skill and aptitude that comes together to effectively determine whether a student is achieving learning goals of a course of study. Stiggins in 1995 described educators who are literate with assessment as “knowing what they are assessing, why they are doing it, how best to assess the skill/knowledge of interest, how to generate good examples of student performance, what can potentially go wrong with the assessment, and how to prevent that from happening. They are also aware of the potential negative consequences of poor, inaccurate assessment." (Reid Lyon & Weiser, 2009, p. 102).

Expertise with assessment also requires a confidence that allows an educator to face a failure of their instruction. Some teachers (and principals) fear that systematic assessment of achievement is risky either because they will find out that the students already know what the teacher is planning to teach them or that the teacher or school failed to teach them properly (Stiggins, 1991). Such failure is inconvenient for planning and doesn’t fit in the typical headlong charge through the curriculum that is characteristic of modern accountability-driven public schools.

Webb (2002) describes the assessment literate educator operating in a standards-based accountability system as one who understands what different assessments do or do not measure, and what can or cannot be concluded from a sample of items on an assessment. Such an educator is not only able to make effective generalizations about student competence but he understands that sometimes assessments have consequences for students, teachers, and staff that create both incentives and disincentives to effective instruction and learning.
Expertise and quality

The answers to the research questions that guide this study are dependent on clear criteria describing different levels of teacher expertise with reading assessments and the relative quality of inferences they make about student learning. Such criteria would also provide a foundation for describing levels of Pedagogical Content Knowledge and quality of instructional decisions.

A variety of researchers have attempted to capture elements of expertise, both as a general topic and specifically referencing reading instruction and assessment. Elements include differences in the way experts solve problems, structure conceptual frameworks, notice pattern anomalies, and make meaning out of ambiguity (Airasian & Jones, 1993; Berliner, 1988, 2001; Reid Lyon & Weiser, 2009).

As a model of identifying expertise as a teacher, the National Board Certification program conducts a rigorous qualification program designed to identify the best teachers in the nation. It has identified patterns of expertise in teachers that address a broad array of skills, attitudes and interests distinguishing expert teachers from the rest, such as a high degree of opportunism and flexibility in their teaching, greater sensitivity to context, and display of more passion for teaching (Berliner, 2001). In a similar vein, the California Standards for the Teaching Profession which were developed jointly by the California Commission on Teacher Credentialing (“California Standards for the Teaching Profession (CSTP),” 2011) and the California Department of Education include descriptors of teacher quality organized into rubrics that have provided a foundation for new teacher support programs (e.g. BTSA) and many teacher evaluation programs at the District level, including my own in the Albany Unified Schools.

Critical work to define a model of expertise in assessment development was completed by the Berkeley Evaluation and Assessment Research Center (BEAR) (Shepard, 2004; “UC Berkeley BEAR Center,” 2012; Wilson, 2009; Wilson & Sloane, 2000). The BEAR system designates four principles of sound educational measurement – 1) a developmental perspective; 2) a match between instruction and assessment; 3) management by instructors to allow appropriate feedback and follow-up instruction; 4) generation of quality evidence. Wilson (2005) describes the BEAR system’s four building blocks of quality assessments that are derived from these principles – 1) a construct map that describes the developmental structure of the knowledge or skill being assessed; 2) a design plan, which is comprised of the questions, tasks, or problems that are used to generate evidence about student progress; 3) an outcome space which is the organization of response data (i.e. categories and scores) as evidence of the construct; and 4) the measurement model, the statistical model used to organize scores into general conclusions.

Assessment Validity

Assessments that are expected to be useful for broader analysis of student achievement have to be seen as valid. Modern validity theory expands previous conceptions of validity with the test construct at its center to include teacher and student perspectives. It is a constructivist process wherein the validity of an assessment is about the appropriateness of interpretation of the data assigned to a student’s performance. In other words, it’s not just the performance task itself, but how the
evaluator interprets the data that determines its validity (Cronbach & Thorndike, 1971; Kane, 2008; Lissitz & Samuelsen, 2007). Validity strength is derived from a well-constructed theoretical framework, grounded in relevant evidence. Investigating validity is akin to scientific inquiry whereby evidence of observable phenomenon can be independently evaluated and subject to replicable procedures to arrive at similar conclusions (Frederiksen & White, 2004; Kane, 2008).

The current system of high-stakes testing for accountability purposes is reliant upon a belief that the tests are accurate and valid indicators of student achievement. These high-stakes tests are supported by a formalized system of validity measures that date from standards developed during the 1940's and 1950's. They rely on a ‘trinity’ view of validity, including content, criterion, and construct validities (Goodwin & Leech, 2003; Lissitz & Samuelsen, 2007). These measures focus on the tests themselves and are independent of use by teachers and learners (Tittle, 1989).

The seminal work of Cronbach and Meehl, “Construct Validity in Psychological Tests” (1955 – as reported in Lissitz & Samuelson, 2007) firmly established “construct” validity as the center of a unified theory of test validity, one that strived to bring a scientific objectivity to educational testing. It provided a solution to the problem of measuring a quality for which the tester has no definite criterion measure. In other words, the quality underlying the test is the important issue, not test behavior or scores, e.g. does a reading test actually test the ability of a child to read, not at what level does she read.

It is this idea of construct validity that is the foundation of the modern ‘high-stakes’ accountability system with its psychometric indicators and applicability to any testing environment or population. More recently, however, modern validity theory has evolved from this post-positivist belief in the inviolability of psychometric calculation to a social constructivist epistemology which supports the idea that validity in assessments is a function of teacher interpretation and related inferences about instruction and student learning (Cronbach, 1988; Kane, 2006; Moss, Girard, & Haniford, 2006).

Reliance on such a narrow, unitary concept of validity has weakened as a broader perspective of validity has developed to include interpretations and uses of a test. Tittle (1989) expanded the notion of validity to include teacher and student perspectives as necessary information in validating test scores, stating that teacher sense-making of assessment is a constructivist process, that “valid inferences can only be constructed by the teacher, and valid use arises from the accuracy and appropriateness of the meaning teachers construct from test scores and the information system in which they are embedded.”

This evolving concept of validity was verified in the 1999 edition of the AERA Standards for Assessment,

"The degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests...The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself." (American Educational Research Association,

Kane (2006), in his standard-setting “Validation” article in the 4th edition of Educational Measurement, described validity as an argument that analyzes the coherence of evidence for or against proposed interpretations, inferences, and uses of assessments. He claimed that arguments about inferences can and should be made at every step of the way from scoring the observed performance by students to the consequential decisions made about a student based on the assessment results.

Moss et al (2006) support this idea by suggesting that evidence of complex systems influencing how learners and the environment interact contributes to the social construction of meaning by teachers, students, and the community.

The accuracy and appropriateness of interpretations and inferences referred to in this updated approach to validity is dependent upon a well-designed and constructed assessment system, such as the BEAR system described in the previous section.

For the purposes of this research, the complexity of the constructs of teacher expertise and Pedagogical Content Knowledge (PCK) required a comprehensive rubric that describes different levels of skill and knowledge in relevant domains. Such a rubric facilitates evaluation of teacher inferences and decisions about instruction (Jonsson & Svingby, 2007; “Process: Why use rubrics?,” 2012). A pair of rubrics was designed as a strategy for understanding changes in sophistication of teacher thinking about assessments and instruction. They are described in the methods section below.

Reading Assessment
An important subset of my knowledge base is reading assessments. Those used for high-stakes purposes are generally acknowledged to be unsatisfactory because they are not domain specific, show low inter-correlations with other reading assessments, narrow the curriculum, and have little value for planning curriculum (Snow, Griffin, & Burns, 2005). Best practice research on reading assessment favors a multi-factored approach that addresses the multiple dimensions of reading, rather than the unilateral approach used in most accountability systems (Marilyn Wilson, Schendel, & Ulman, 1992).

To truly develop a thorough understanding of a students’ reading proficiency, the multi-dimensional nature of the content knowledge for reading necessitates a comprehensive and complex set of assessments that address each of the major reading domains, including comprehension, fluency, decoding, and text structure.

The centrality of reading comprehension in the acquisition of knowledge makes its assessment a most important endeavor for preparing students for success in higher education and the workplace. Assessment of comprehension has been a particularly vexing methodological problem because of its complexity as a domain and the fact that so much of reading comprehension is cognitive and can’t be seen as it happens (Valencia, Pearson, & Wixson, 2011).

The research described here examines the history of how the domain of reading comprehension and its assessment developed (D. Pearson & Hamm, 2005; Hiebert & Raphael, 1996), and how the domain is defined developmentally (Alexander, 2003). A
model for development of complex “through-course” assessments is described and evaluated (O’Reilly & Sheehan, 2009; Valencia et al., 2011).

Traditionally, taxpayers, legislators, parents, and district administrators were rarely privy to the results of assessments that have taken place daily in classrooms across the world, such as written notes, interactions during discussions, and scores on comprehension quizzes. From the 17th to the 19th centuries, reading skill was measured through accuracy and fluency in oral capacity in the tradition of “declamation and oratory” (D. Pearson & Hamm, 2005).

Hiebert and Raphael (1996) documented a changing accountability environment that was based on a new field of educational psychology. Thorndike’s 1909 presentation of a handwriting scale is largely credited with initiating this new era of a “scientific” measurement of learning. Industrialization in the early 20th Century, associated demographic movements and an increase in school enrollments of students with poor literacy skills increased the need for cheap tools to determine student levels of literacy. The development of intelligence tests that included reading comprehension measures, and then the introduction in the mid-1930’s of an IBM scanner that enabled machine-scorable multiple choice tests began to transform the testing environment as did the development of factor analysis as a new statistical tool in 1944 (D. Pearson & Hamm, 2005).

Recently, however, research on modern reading assessment and its typical treatment of prior knowledge and the nature of passages and questions, has exposed significant flaws in the standard multiple-choice testing approach that has dominated the past 50 years of reading assessment for accountability. Tests are designed to “wash-out” the effects of prior knowledge because passages on which majority of students have a lot of background knowledge don’t remain on tests because everyone gets high scores; and the typical short paragraphs that are the basis for testing results “often lack structural and topical integrity, making it difficult to create questions that tap higher level literacy processes such as inferencing” (Hiebert & Raphael, 1996, p. 565).

In modeling a new approach to reading assessment, Alexander (2003) proposes an approach that recognizes the interaction between relevant cognitive, motivational, and sociocultural forces. Her Model of Domain Learning (MDL), which describes a recognized field of study with its own community of practitioners, has a particular lexicon, established routines and rituals. For reading, it includes subject-matter knowledge, learner interest, and strategic processing with three stages on the developmental continuum toward proficiency – acclimation, competence, and proficiency/expertise.

This developmental approach to reading instruction and assessment is best served by “through-course assessments” that take place closer in time to when key skills and concepts are taught, making the information more actionable. O’Reilly and Sheehan (2009) detail a new developmental model of domain competency – the Cognitively Based Assessment of, for, and as Learning (CBAL) which is used as a basis for the design of instruction, accountability, assessment and professional development.
Cross-Case Study/Action Research

I used an exploratory Cross-Case Study approach within an Action Research framework. This means that I examined the experiences of several different teachers as they engaged in a change-oriented reading assessment experiment to help resolve real problems they are facing within their school (Coghlan & Brannick, 2005).

A case study is appropriate in this situation because the research question asks how something is done in a complex environment with many variables that affect the answer to the question. Case study research includes “detailed, in-depth data collection involving multiple sources of information (e.g. observations, interviews, audiovisual material, and documents and reports)” (Cresswell, 2007, p. 72).

An Action Research approach is called for because the implementation of a complex reading assessment plan is a collaborative process that “aims at both taking action and creating knowledge or theory about that action” (Coghlan & Brannick, 2005, p. xii). The research itself is an intervention, as much of action research tends to be, that “typically involves reeducation, a term that refers to changing patterns of thinking and action that are currently well established in individuals and groups” (2005, p. 10).

The epistemological assumption of the action researcher is that the purpose of academic research is not just to describe and understand the world but, in fact, to change it. In some ways, the very act of framing and selecting an action research project, such as that which is described in this paper, is an important step in the change process itself.

Conducting action research when the researcher is a manager in the subject organization requires nimble footwork to maintain trust of the other participants and the organization as a whole. It requires situating the project in the organization’s learning-in-action culture and carefully managing role duality issues and access to resources. For example, I play several relevant roles that may have some impact on the outcomes of the study, including my district leadership role as Principal of the high school in which part of the research occurs, my trainer role, and my role as a doctoral student which means that the results of the study have external reporting implications.

I have helped develop the learning culture of my district and can access most anyone in the district at any time. Access, however, can be a double-edged sword. It creates concerns of objectivity in terms of the researcher’s use of appropriate frameworks for viewing and understanding data. It also raises questions about the research write-up, feedback given to colleagues, and how all of the participants handle negatively perceived interpretations.

A critical approach to the answer of these questions is the regular journaling by the researcher, a reflection process that focuses particularly on what Coghlan and Brannick call “role duality” (2005, p. 68). There are many elements of the research process that can be affected by my professional relationship with the case participants. The journaling process helps me explore the lens through which I view and understand the data, how the research may be written up, and how I might handle negative interpretations.
Research Design

The research design has several key components, including quantitative measures of teachers’ general assessment literacy (Assessment Literacy Inventory) and their self-reported metacognitive reading processes (MARS – Metacognitive Awareness of Reading Strategies Inventory), which helps with evaluating their Pedagogical Content Knowledge of reading. Most of the data, however, was collected through a series of structured interviews designed to delve into a teacher’s thinking about student skill in relation to the specific assessments; and relatively unstructured “think-alouds” where they discussed their thoughts about how they could, would, or did use the assessment results in their ensuing instruction.

At the student level, reading assessments are an essential part of the design. Reading assessments administered to students included a lexile-based test to measure a student’s technical reading skill and comprehension, the Test of Silent Contextual Reading Fluency (TOSCRF) and Test of Silent Word Reading Fluency (TOSWRF) to measure silent reading fluency, context-based vocabulary assessments to measure word knowledge, and the MARS to assess metacognitive processes. These measures provided data to correlate with teacher thinking over time.

The design included some level of teacher training in the use of the assessments. I conducted the teacher trainings at a basic level between the first and second student assessment administrations and monitored their evolving skill and understanding through the periodic interviews. The training focused on Alexander’s “Elements of Development” (Alexander, 2003), O’Reilly’s “Classes of Reading Skill” and “Cognitive Skills Framework” (O’Reilly & Sheehan, 2009), and Valencia’s “Habits of Good Readers” (Valencia et al., 2011). The training essentially consisted of the handouts (with citations) in Appendices 7 and 8 and a discussion of each. The high school teachers were brought together and trained as a group, while the middle school teachers received the training one-to-one.

Case Selection

Volunteer teachers from the Albany Middle School and Albany High School English Departments are members of the bounded system for the case study. At Albany High School, as part of a District-wide initiative to align instruction and assessment to a set of synthesized content standards (essential standards), the English Department decided to pilot a complex series of reading assessments to measure and monitor the development of reading skill among the middle school’s students. At Albany Middle School, two teachers agreed to participate – an English/History Core teacher and the Reading Intervention specialist.

Both schools are high performing schools in a diverse, middle class community just north of the City of Berkeley. However, the middle school was concurrently in its first year of Program Improvement (PI) status due to one subgroup’s failure to achieve

---

1 Vocabulary test items designed according to the “Degrees of Contextual Reliance” strategies outlined in Pearson et al (2007, p. 289).
its prescribed Annual Yearly Progress (AYP). Demographically, the school is 52% white, 31% Asian, 8% African-American and 2% Hispanic/Latino. 12% of the students are English Learners, and 21% are lower SES. The high school is more diverse demographically with 35% white, 40% Asian, 7% African-American, and 15% Hispanic/Latino. 13% of the students are English Learners and 21% are lower SES.

This research falls within the category of what Coghlan calls “mechanistic-oriented” action research. I was not studying my own actions within the system; rather I was studying the system itself. The purpose (as described in the introduction) was to get at, or confront, a pre-identified issue, that of the teacher’s use of assessment information to make instructional decisions (2005).

**Data collection**

As stated previously, I used focused interviews of individual teachers, pre and post surveys about assessment literacy (Assessment Literacy Inventory - ALI), the MARSI (Metacognitive Awareness of Reading Skills Inventory), and think-aloud protocols about the grading process (Orrell, 2008). The ALI is a psychometrically vetted assessment of teacher assessment literacy ((Mertler & Campbell, 2005, p. 12) although the authors of the assessment acknowledge the need for further analysis of its validity and reliability. The MARSI, on the other hand, makes no pretentions about being a measure of a reader’s “comprehension monitoring capabilities”. Rather it is designed as a tool to help students increased their own metacognitive awareness and strategy use while reading (Guan, Roehrig, Mason, & Meng, 2011; Mokhtari & Reichard, 2002, p. 255).

The previously described rubrics about assessment literacy and Pedagogical Content Knowledge provided valuable data about the expertise of the teachers and the quality of their inferences and decision-making. These approaches helped to inform the theory of action described in Chapter 2 by triangulating the varying forms of evidence collected. The quantitatively oriented measures, like the ALI and the rubric-based scoring of interview excerpts, provide a psychometrically vetted source of information that is balanced by the self-reflection of the MARSI, the think-aloud and interview protocols.

Open-ended interviews are an opportunity for key informants to provide insights, corroboratory evidence and suggestions about different research threads to pursue. Focused interviews are still open-ended but follow a set questions derived from the knowledge base (Cresswell, 2007). On the other hand, structured interviews, as compared to open-ended interviews, present an opportunity to more formally and reliably aggregate the data because the same questions are asked in the same order and the interview protocols are pre-designed.

I introduced the reading assessments and research plan in the last part of the Fall semester, 2011, secured participation commitments from teachers and began implementation shortly after the beginning of the Spring Semester, 2012. Only the data generated from teacher interviews, the ALI, and the rubrics were used in my analysis. Student data from reading tests were used solely as indicators of student reading skill as a source of information for teacher inference and interpretation. The student data is
not being reported as part of this research in any way, although individual student data may have been reported to the student and parents per teacher discretion.

My data generation plan followed the order of operation as illustrated in Figure 4 below, including a series of three focused interviews of individual teachers (see the interview questions in Appendix 1), the pre and post surveys about assessment literacy (ALI – Appendix 3), and teachers’ metacognitive reading strategies (MARSII – Appendix 4). I kept track of individual teacher participation in the various aspects of the study through the Record of Participation (Appendix 2).

![Figure 4: Order of Operation](image)

**Data Design**

To answer my research questions and probe the theory of action described previously, I had to be able to capture data from a variety of sources about different aspects of teacher knowledge, thinking processes, and decision-making. Case studies typically rely upon a process of triangulation to guide the analysis of multiple sources of converging data (Yin, 1994), thus, I used a mix of quantitative and qualitative data to get at the issues.
The participating teachers took three assessments that helped gauge their assessment literacy and PCK for reading. Pre and post usage of the Assessment Literacy Inventory (ALI) and the Metacognitive Awareness of Reading Skills Inventory (MARSI) provided some quantitative data while taking the TOSCRF/TOSWRF provided an experience with a fluency test that provided useful qualitative data for the 1st interview. My original plan was for students to take a series of 4 different reading assessments at two different points in the research timeline – shortly after the first teacher interview and then after professional development sessions. As the reality unfolded, it wasn’t possible for all of the students to take all of the assessments a second time but they all did take at least two of the assessments twice which was useful to the interviews. Coding from the four interviews with teachers and their “think-alouds” about instructional decisions provided rich information that I was able to statistically compare across measures and at different stages in a teacher’s development.

For the purposes of this research, two different rubrics were developed based on the teacher expertise and assessment quality research described in Chapter 2. The first was designed to evaluate teacher PCK and quality of instructional decisions (Table A); and the second to evaluate a teacher’s assessment literacy and inference quality (Table B).

### Table A:

**Pedagogical Content Knowledge and Instructional Decisions**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>Simultaneously integrates comprehensive PCK (subject-area domain knowledge, contextual factors, and teaching and learning theory) with assessment constructs and outcomes to evaluate student developmental progress and design differentiated interventions</td>
</tr>
<tr>
<td>Proficient</td>
<td>Fluidly recognizes meaningful patterns within the PCK framework to interpret assessment outcomes and prepare targeted instructional interventions</td>
</tr>
<tr>
<td>Developing</td>
<td>Selects from elements of PCK to describe assessment outcomes and review student progress. Instructional decisions reflect assessment outcomes</td>
</tr>
<tr>
<td>Novice</td>
<td>Identifies and labels assessment outcomes and is able to identify high and low achievers within a limited definition. Assessment outcomes have little relevance to instructional design.</td>
</tr>
</tbody>
</table>

### Table B:

**Assessment Literacy and Inference Quality**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>Inferences about student learning are contextualized developmentally and are derived from assessment evidence based on a multifaceted analysis of the domain’s construct, the assessment task design, organization of data, and statistical models used to evaluate student outcomes.</td>
</tr>
<tr>
<td>Proficient</td>
<td>Inferences derived from assessment evidence are based on a rudimentary understanding of how the assessment was designed and the relevance of the results to the domain’s construct. There is some consideration of the data organization and statistical models used to produce results.</td>
</tr>
<tr>
<td>Developing</td>
<td>Inferences derived from assessment evidence are based on rules and protocols provided by assessment designers. There is little awareness of the direct</td>
</tr>
</tbody>
</table>
relevance of a particular assessment item to the construct of the domain.

| Novice | Few, if any, inferences are actually made as a result of assessments. Data is used to summarize the results of student performance. |

Teacher interviews were coded according to these rubrics as one strategy for indicating teacher expertise in these areas. These data were compared to the two administrations of the Assessment Literacy Inventory and any changes in indicators of expertise are noted.

**Data Analysis**

The data analysis identified common patterns and logical threads among the different sources of data. As illustrated in Figure 1, I expected to identify high/low assessment literacy among the participants and high/low pedagogical content knowledge as the predictors for quality of inferences and decision-making for instruction. The results are discussed in Chapter 4.

Interview coding was structured by the conceptual framework described previously and focused on critical incidents and information that elicited the thinking of the participants. Codes were derived mainly from the terminology of the Assessment Literacy Inventory (ALI) and MARSI instruments as well as the training and knowledge base I provided about the various reading assessments to be administered. In Appendix 6, the coding configurations structured around the knowledge base for Pedagogical Content Knowledge for Reading and assessment literacy are identified. A software program, Dedoose was used to analyze coded data from the interviews (Dedoose, 2012). I used StatPlus statistical software (StatPlus:mac, 2012) in combination with Excel to bring the quantitative data from the ALI together with the rubric scores derived from participant interviews to facilitate the identification of patterns. Through analysis of the various patterns and tendencies, I was able to explore the answers to the two research questions presented earlier.

The Data Analysis Spiral, described by Cresswell (2007), provided a useful model for the collection, management, analysis, and representation of the data. Initially, I developed a short list of 5 or 6 coding categories, derived from patterns and categories identified through interviews, observations and the guiding conceptual framework. There were 4-6 sub-categories developed for a combined coding list of roughly 20 categories. However, the analysis was directed to reduce them back to a more manageable set of 5 themes that I used on a dichotomous basis to code interview excerpts (the theme was either present or absent) and to write the narrative (see Appendix 6). They are:

- **Pedagogical knowledge** – including direct and indirect instruction and student engagement. Interview excerpts that described specific instructional strategies or pedagogical approaches, such as read or think-alouds, research, jigsaw cooperative learning methods, multiple intelligences, etc. triggered a coding response.
- **Contextual understanding** – such as the purpose for reading, student background, and the learning environment itself. The purpose of these codes was to help describe the context in which the instruction and assessments took
place providing more facets for teacher analysis and interpretation of student results.

- Reading domain – describing the major elements of the reading domain – subject-area knowledge, strategies, interest and motivation, and content knowledge. Issues of background knowledge, cognitive and metacognitive strategies, factors that trigger reader interest, and the teacher’s knowledge of the key elements of reading e.g. comprehension, vocabulary, fluency.

- Outcomes – which address a teacher’s technical knowledge with regards to producing assessments, particularly how skilled they are at test item design, procedures for collecting data and developing appropriate methods for evaluating the results.

- Assessment literacy – refers to the consumption of information derived from assessments and how a teacher evaluates the assessments themselves. Psychometric issues of validity and reliability are of concern here, but so are the factors that affect teacher inferences about student results. Teacher knowledge about different categories of assessments is included as is their attitude about assessment, i.e. their own motivation and the value they place on assessing student understanding.

My theory of action would be confirmed if the terminology-use and conceptual patterns illustrate a growing sophistication with the reading assessment instruments and I see a concurrent increase in the depth and extent of teacher thinking about student learning. By the same token, my hypothesis would be confirmed if, over time, teachers demonstrated a more sophisticated use of terms and concepts from the PCK of reading instruction combined with data patterns showing how instructional decisions, such as interventions, reteaching, or differentiated learning structures, emerged from higher level inferences about student reading.

Research Validity and Reliability

Validity concerns abound given the nature of the data collection process and the complexity of the research topics. In action research, the researcher needs to create protective mechanisms to keep him honest. Through journaling, I attempted to confront disconfirming researcher bias up front so that the reader is able to understand my biases or assumptions that impact the nature of the research protocols or analysis.

Triangulation of the data provided corroborating evidence from multiple sources and methods. I also offered what Cresswell calls “member checking” which is the process of the researcher asking the participants to review and verify the credibility of the findings and interpretations of the data (Cresswell, 2007, p. 208). Several of the participants took me up on my offer. Their comments and suggestions were incorporated in the findings.

An important test for judging the quality of a research design is reliability, or whether the fundamental operations of a study, such as data collection efforts, can be repeated with similar results. I used the three different strategies Yin suggests for increasing the reliability of a case study (1994, p. 37):
1) A “case study protocol” which contains the instruments to be used in the study (surveys, coding procedures, etc.) and the procedures and general rules that should be followed.

2) A case study database that includes the data collected and the report of the researcher.

3) Finally, a chain of evidence was maintained. This chain allows an external observer to follow the development of evidence and analysis from the first research questions to the final case study conclusions.

To verify the reliability of the scoring, I sent 28 excerpts that I had scored according the two rubrics represented in Tables A and B along with copies of the rubrics to two educational professionals and fellow students in the Graduate School of Education at UC Berkeley. I asked each of them to score ½ of the excerpts. They did not receive any training, nor were there anchor papers or any kind of calibration. To be able to calibrate the evaluations effectively, significant training in the domains and the nature of the rubric would be required.

Despite the lack of training, the consistency of the scoring was remarkable. On a four-point rubric, 9 of the excerpts were scored the same, 16 were different by one, and 3 were different by two points. In traditional holistic scoring, one point is considered a normal difference, so by this measure, the rubric was scored either identically or within a normal variation on 90% of the excerpts. That’s pretty good.
Chapter 4: Findings

I expected to find that teachers who are stronger in their pedagogical content knowledge and score higher on the assessment literacy criteria are more likely to model their decisions about instruction based on the data. This means that when the data indicates a student has not learned a concept or skill, the teacher will be more likely to make an intervention of some sort.

As teachers gain more experience with an assessment and spend the time to reflect on which aspect of the domain it actually measures, then they are likely to make more accurate and useful inferences about what the student has learned and what he needs to trigger deeper and broader learning. I expected that given this experience and reflection and an expanding PCK, that the instructional decisions made would be of a higher quality as they would reflect greater understanding of the learning needs of the students in relation to the content.

The findings section is divided into three parts – an examination of the developmental experience of each of the 6 cases; general patterns identified and illustrated through the comments shared in the interviews; an analysis of the statistical data generated; and a synthesis of the results.

Profiles of Cases

Each of the 6 cases profiled below illustrate the complexity of the issues I investigated. The period of data collection extended across a single semester which wasn’t enough time to see major changes in the thinking and instructional behavior of teachers, some of whom had been teaching for decades. Yet, it was exciting to witness the ballooning perspectives of these six English teachers (only one of whom, Betsy, was specifically designated as a reading teacher) as they experienced the variety of reading assessments themselves, reflected on what they measured, administered them to their students, analyzed the results in a guided fashion, received professional development on the reading domain and assessment, administered the assessments a second time, and then reflected on the results and what they mean for their instruction.

By the third interview, each of the teachers were more fluent in their analysis of the results, described the depth and richness of the multitude of factors that are at play both in student performance as well as in their interpretation of the results, and were motivated to teach reading in fundamentally different ways than they had previously.

Sandy\textsuperscript{2}. After three years in the teaching profession, Sandy’s passion for developing as a teacher was increasing. While involved in this research, she was also enrolled in a Masters Program at a local university. She is a member of two different small learning communities at the high school so is a highly collaborative and organized individual. Her relatively low results on the ALI, while they did go up on the second administration, are evidence of her lack of training and experience.

From the beginning of our conversations she was engaged and digging into her practice, stimulated by the various measures she took. By her own accounting, she

\textsuperscript{2} All of the names are fictitious.
was an enthusiastic proponent of assessment, both the formal approach as well as the informal little indicators of student understanding such as “exit tickets” or thumbs up/thumbs down. She sees teaching as “a series of academic based stepping stones and life skill-based stepping stones, both within a class period, a unit, and within the school year”. Assessment helps the teacher “identify the next stone to step on”.

The ALI, in particular, surprised her with the way it presented the “teaching experience distilled down to a multiple choice format” and caused her to reflect on her own assessment practices. Writing instruction is a particular emphasis at the school and she thought about how difficult it is to quantify the writing process.

Writing is so subjective from both the writer’s perspective and the reader’s perspective. To what extent, I mean I can represent an individual’s ability to write with numbers and how can we standardize the essay grading process because we can’t do a multiple-choice test for an essay. I mean we can get at different skills within an essay with multiple choice but the magic is done when an essay is polished. It’s so hard to quantify.

Sandy’s perspective on the MARSI evolved rapidly during the period of the study. She gave herself a very high MARSI score (96%) but then she dismissed it saying that none of her students would be able to understand the words. Despite this initial skepticism, and after she administered the assessment to her students for the first time, she was able to fully consider its value in relation to the other measures. She described its “incredible” value because of its ability to take “this enigma that is the reading experience and it divided it into 30 concrete tasks” many of which could then be divided into discrete lessons. She talked about it becoming a high priority as a teaching tool, a centerpiece of her instructional strategies.

(Reading instruction) needs to be really, really, highly scaffolded and differentiated. The metacognitive layer… I think it’s the easiest to weave in. I think it’s always a really interesting conversation for the students because they just love to talk about themselves.

After the second round of assessments, when she was able to think about the realities of differentiating, she noted how difficult it is to meet the needs of students in a typical classroom.

I wish that I had was just more time to work one on one… I mean, trying to teach 30 different reading levels at once is very, very challenging.

The typical classroom teacher has so little time to sit and think about a particular student and to consider what the results of a specific assessment mean. This is one of the biggest problems with teachers making valid inferences about student results. They don’t have the time or training to really try to figure out what is going on with a student and their development in a skill area. Sandy’s experience with the 1st administration of the silent reading fluency measures, the TOSCRF and the TOSWRF were good examples of this. As much as she wanted to understand the implications of the results,
all she was really able to do was come up with some plausible ideas, recognizing that she was shooting in the dark, really just guessing.

I feel like, if anything, this (the TOSCRF scores) might be an indicator of just experience… even if it’s not telling me their ability to understand it, like how frequently they move their eyes over words. And from that, I might be able to think about how confident they feel reading, or how much they enjoy reading, or how much they would sort of feel intimidated by a piece of reading. But those, I felt like I was stretching it a little bit, like I would definitely want to piece this piece of data with other experiences that I’ve had with the students in different kinds of texts.

The vocabulary assessment took an intriguing turn by the second administration because I spent quite a bit of time with Sandy reviewing the types of questions that had been incorporated into the test and we had debated a bit about which of the 4 types of questions she wanted to use for one particular term, ‘comedy’. This term was an excellent example of how difficult it can be to create test items that match a particular design. The item is reproduced below:

That play is categorized as a comedy because it focuses on errors of mistaken identity and how characters are able to solve them.

a. Humor
b. Tragedy
c. Farce
d. Drama

The term was introduced in the context of her class’s study of Romeo and Juliet. She became frustrated by the question because the item design called for distractors that were generally less common or related meanings of the word, but they were perhaps too unfamiliar to the students (9th graders).

They reflect a much more sophisticated and nuanced understanding. So that was really challenging for me because it was asking them to know the subtleties of these different really sophisticated ideas when we had just been dealing with the concept in a generalized way.

This example illustrates the challenge both of designing assessment items that are useful in the context of a single teacher’s classroom and those that are useful as a commonly administered benchmark or standardized test for accountability purposes. For a student to be able to answer the question above, their ability to distinguish between the somewhat unfamiliar theatrical terms becomes key to their success with the question. The term ‘comedy’ is familiar to all of them but when tested in a theatrical context with other more academic terms, especially when the definition is the salient context, then a precise understanding of the meaning becomes more important. Given the number of ways to teach the meaning of the word comedy and the varying levels of complexity of the term, the instructional component becomes very complicated as does the ability of any one question to measure a student’s understanding of the word.
In Sandy’s case, for the second administration of the vocabulary assessment, she had actively and directly taught the vocabulary being tested (as opposed to the first administration which was a pre-test). She noted that the meaning of the words that was taught to the students didn’t match closely with the correct answers on the test which served to confuse the students a bit.

*Those were words that we really looked at and analyzed how Shakespeare was using them. I taught them completely by the time we took the test. This was a posttest, and they thought that I was really looking for those specific meanings, so that sort of threw them off.*

By the third interview, Sandy’s growing PCK had an impact on how she looked at assessments. She looked askance at multiple-choice tests, concerned about their limitations and thought about different strategies that might capture the complexity of the reading construct.

*I think there’s something inherently really, really hard about analyzing something that complex through a multiple choice test…I think ideally, if I were really to assess knowledge of a word, I would want them to create something like a word web or something so they could—they were generating the different meanings and to show their understanding of the word that way.*

The increase in the sophistication of her comments about the reading domain, assessment results, and student learning did indicate that she became fully engaged as the study progressed even if she may have been a bit reluctant in the beginning. Her relatively low initial ALI of 69% made a small gain to 72% by the second administration.

**Greg.** Teaching requires a fairly high degree of confidence on the part of the teacher to be able to spend the hours in front of a classroom directing students through the learning process. That confidence is typically well-founded within an expertise in a particular narrow subset of a subject-area but sometimes, a teacher feels like he has to demonstrate the same confidence for each subset of the subject that is being taught in order to maintain his academic authority.

Greg is an individual with a bit more experience than some of the others, and he proved to be astounding knowledge and bright on some topics but stumbled on others. Despite 7 years of teaching under his belt, Greg had among the lowest of the ALI scores (he felt he was being assessed more on his familiarity with the language rather than the content) and had experienced very little professional development specific to reading instruction. While he generally felt that assessment is an important part of teaching - “otherwise you’d just be some narcissist talking to yourself” - it took a while to break through his veneer of confidence to get him to recognize that he really did have a lot to learn and he was ready to be engaged. Early in the process, he associated assessments with tests and quizzes, and commented this was not his favorite part of a teacher’s work in part because he felt assessments could not capture enough of the student’s skill development.
Grading students is not my favorite part of being a teacher and I associate assessment with grades and judging...To kids, it must feel like there’s some form of judgment being passed on them and their skills. Kids that I have to give really low grades to, and kids who have come a really long way, sometimes it’s hard to communicate that in a formal assessment they’re still behind.

His initial comments about what makes a teacher good at assessments did not address elements of validity. Nor did he say much about how an assessment is designed. Rather, he focused on using a variety of measures to enable a student to demonstrate their knowledge or skill, a decidedly multiple intelligences approach.

I think a good assessor is mixing it up, is providing assessments that aren’t just essays, but they’re providing speeches and other ways for kids to demonstrate their intelligences and that measure sometimes incorporating growth and sometimes based on skill.

After the first administration of the assessments to students, Greg tried to make sense of the different kind of information each provided. As he brainstormed ideas about what they meant, he identified various factors that would have to be evaluated as part of the context for teacher interpretation of results. For example, as he talked about the MARSI, an initial comment that it is a “fairly accurate way” to determine a student’s reading level was followed immediately by thinking about the limitations of the measure. The MARSI being reflective is... I think trying to assess a kid’s awareness of their reading strategies, which is a little bit complicated because whether you’re aware of them or not, you may not have them reading strategies wise. And just because you’re aware of them doesn’t mean you’re actually employing them effectively.

A similar thing occurred as we discussed the vocabulary tests. He instinctually was aware that it wasn’t just memory at work with student scores. Students had to engage at a deeper analytical level to figure out the answers to the questions. It’s definitely an assessment of their inference making ability. Their ability to contextualize and... to recall past stories, past conversations where they’ve read or heard it, and then obviously just the context of the sentences.

Examining individual student results, he was fairly perplexed at some of the inconsistent results that were produced. He described one student as a girl who really struggles in the class and is not getting good grades, but her scores on the standardized tests, the vocabulary, and the reading fluency were pretty high. This to me, I have no idea, but... I don’t know how to interpret the data, really.

Yet, when it was pointed out that her MARSI was relatively low, his inference about her skills and instructional needs began to develop, beginning to realize the role of metacognitive strategies in reading comprehension.
It seems like she has the natural, not natural, but she has one full set of the skills and yet is missing something that would allow her to better apply them.

By the third interview, after the 2nd administration of the reading assessments, Greg was increasingly reflective about his own reading practice and was using this knowledge to frame his instruction, actually describing his metacognitive strategies to students and identifying which he scored low in and thus didn’t use instructionally. This was part of an effort to help them examine their own use of the strategies.

He made an interesting connection between the cognitive strategies, particularly visualizing, and the use of strong visual descriptors by some authors, commenting on how different students who rely on different strategies respond. Relevant to his knowledge of the students and the context for learning they provide, he noted that some students are drawn to the writing of John Steinbeck, who is famous for the imagery he conjures through his writing, while others were impatient with such detailed setting of the scene.

In 9th grade they read Steinbeck and, you know, you can’t not get a visual out of reading an opening chapter and a new scene in his writing. And it’s interesting because some of the kids process that as really good and friendly for them in reading, and other kids process it as dragging on and wishing he would get to it.

Of the different elements in the reading domain that we had assessed, Greg was clearly most interested in the MARSI and teaching the metacognitive strategies. Vocabulary didn’t really grab him as much. When asked about his instructional strategies, he relied on a traditional approach including writing down the dictionary definition, coming up with a few synonyms, discussing the part of speech, the pronunciation and then they are assigned to write sentences using the word. For the visual learners, they might play Pictionary. It is geared to memorizing a word and experiencing it in different contexts.

He remained somewhat perplexed about how to teaching reading fluency, not really thinking about its role as an indicator of challenges students face nor as a construct that has many factors contributing to it, such as background knowledge, language familiarity, visual processing issues, motivation, etc. As he tended to do, he personalized his inferences, attempting to understand the student experience through his own lens.

We were talking about picking up the actual fluency, the pace of reading. And I don’t really know how to instruct that at all. As an English teacher, I’m like well, it’s not math or science, you’re supposed to read all the words, you’re supposed to try to enjoy it a little bit.

Thinking about changing the way he assesses reading for the future, he was clearly most interested in the MARSI and its usefulness as an instructional tool. Surprisingly, after some reflection, he decided that he also wanted to improve his skills with teaching and assessing student vocabulary. He recognized that his methods were
a bit outdated and he wanted to develop his skills with multiple-choice tests as one assessment strategy.

*I never really understood how to do a multiple-choice test. It always just seemed like if I made one it would just end up so obvious what the answers are. So considering… And maybe like practicing a little bit… at least trying to construct my own vocabulary test that way is definitely one thing I would do.*

**Betsy.** A 10 year veteran, and a very engaged middle school teacher, Betsy has a Masters degree and has participated in a fair amount of professional development focused on reading. She teaches a reading intervention class along with her regular teaching duties. This increased her awareness and motivation to learn more about the reading domain and instruction. It also provided her with more opportunities to experiment with various assessments and instructional strategies as she was looking for ways to meet the needs of the students, particularly in the intervention class.

She took a broad perspective initially about assessment, as do many teachers, including the most informal such as “*even just a verbal fist, thumbs up, thumbs aside*”, but also recognized that she wasn’t the most knowledgeable about reading assessments and one of the reasons she was participating in the study was to learn more, “*I don’t feel very confident about what kinds of assessments are out there.*”

Betsy’s ALI started at a mid-range 74% correct and didn’t change at all between the two administrations. This was a bit surprising given the amount of time we spent discussing issues relevant to assessment literacy and the fact that the other teachers who took the ALI twice increased their scores, even if only through familiarity with the assessment. Her MARSI was on par with most of the rest of the teachers (except for Sandy as mentioned above).

The MARSI probably had the biggest impact on her thinking and practice, in large part because it made her reflect on all of the different strategies, which ones she used effectively and which she didn’t, noting that there was a connection between her own struggles as a reader and the strategies she didn’t use effectively.

*I noticed that the text that’s always been more difficult for me has been - since I was in school - expository texts, and that some of these strategies that I don’t use when reading those kinds of texts such as ‘using text features and structure to understand expository texts’ are probably the reasons that it was a type of text that was hard for me to approach…This triggered in my mind: Maybe that’s why I couldn’t understand my history or maybe that’s why science textbooks were really always hard for me. And to this day, reading academic articles… Now I try to do these things just because I know more. It did help for me as a mature reader to reflect on my learning in the past, and to see why I do well and why I don’t.*

She also made a comment about how it takes a certain amount of “confidence” for a teacher to be good at designing assessments, interpreting them accurately, and then making decisions about what to do with that information.
I think what makes a teacher good at designing assessments is, again, a real confidence in what the underlying goals are for the lesson, for the curriculum, and a good knowledge of their students' abilities and level of development...I don't think enough of us really feel confident about the goal that they're trying to teach. I think it's very blurry. Let's take a piece of literature. We know what we want students to come out with generally. There are a lot of things that go into teaching literature. But how confident do you feel about prioritizing? What is the most important thing... what skill, what knowledge, what facts... How do you prioritize what really is the goal of teaching that? I think we don't feel confident.

As students took the different assessments, Betsy's understandable lack of knowledge about the design of the assessments and what they were trying to measure became clear. For example, her perspective on the vocabulary quiz after the first administration was that she blew it, “The results tell me that I didn’t teach very well, too. I didn’t teach those words very well.” There was no thought about looking at the actual questions to try to understand what they were measuring to determine whether they even matched up to what she had taught.

In another example, she examined the results of the first round of assessments and noted that one student, whom she described as a “fluent reader” had scored high on several of the assessments but low on the MARS I. At first, she suggested that he was just not taking it seriously but when I reminded her of the different types of readers we had learned about during our professional development session, she realized that his issues might be connected to his background knowledge and not to his use of strategies, thus he might be “fluent” with lower level texts, but he might not have the strategies to facilitate his comprehension of more complex texts. She wasn’t willing to give up the idea that he disengaged for the MARS I.

The amount that the student reads means that there is a tremendous vocabulary back there that’s just swelling up. I think, also, personality-wise, doesn’t want to recognize something that he might be using.

She was a bit surprised to see that the assessment results were generally consistent with each other, that a student’s strengths in reading were fairly reliably tested across the different measures.

What grabbed my attention was that they, for the most part, were nice verifications of each other. That the tests verify what the other tests had shown. In other words, if a student shows a pretty high CST score and a pretty high Gates-MacGinitie, these other scores, for the most part, also said yes they used strategies, and yes they’re pretty fluent, and yes while they’re reading quickly, they’re retaining memory of what they’ve read and they’re putting some comprehension into it with the context fluency. So it was nice to see.
On the other hand, in one particular case she found inconsistent results - an ELD student who scored high on vocabulary and metacognitive strategies but low on the other reading measures. She described him as a “careful, plodding, hard-working student.” But when I suggested that perhaps his effective use of strategies and plodding approach made up for his lack of background knowledge and fluency with English, she agreed. “Maybe that isn’t inconsistent, you’re right. And that makes sense.”

By the last interview, Betsy was beginning to zero in on changes she had experienced and different ways she wanted to reorganize her instruction. Learning about the different vocabulary question types and thinking deeply about how they represent student vocabulary knowledge, led her to want to fundamentally change how she dealt with vocabulary.

*That article you gave me that broke down the vocabulary levels, ways of understanding the words was really new to me. Let’s put it this way: it changed my method of teaching... It did really make me think about, ‘oh, let’s not just teach it in the way it’s taught for that chapter in this one text. Let’s think about where they would see it and use it in other contexts.’*

She was very interested in exploring the role of reading fluency in her instruction as well, trying to figure out how to teach students to be able to increase their silent reading fluency, a topic we did not address very well in the professional development portion of the study.

**Albert.** Each interview with Albert lasted longer than those with any of the other teachers. His engagement with the research process and his own growth as an educator was a dominant quality threading through the interviews. It was a struggle at times to keep my researcher hat on straight as our conversations kept trending towards a more collegial and intellectual discussion about the topics raised. Albert was the most experienced of the teachers with 23 years as a teacher, many of those at the college level. He has a Masters degree and had experienced a fair amount of professional development about reading at various points in his career although, like most of the English teachers at the high school, the preponderance of his collaborative professional development has been with the teaching and holistic scoring of writing. Most notably, his ALI started out very high at 89% and grew from there to 94% by the second administration. Despite the high score, he commented on the relatively unfamiliar language in the ALI assessment, which raised questions about its validity as a measurement of assessment literacy.

*Albert: It was actually difficult because thinking about the language you use as an educator with testing and assessments is rarely shared enough among teachers and we don’t use the language a lot. And you don’t use it with students. So using it, reading it seeing it, thinking about it. It was, I thought, very challenging.*
As might be expected of someone with such complete engagement in the process, his commentary on student assessment results started out with a high degree of engagement attempting to figure things out but didn’t reflect a lot of sophistication with the domain. The example below is a response to my question about the importance of assessment, but it illustrates his difficulty with developing coherent thoughts about the topic, especially in an open-ended interview process.

Albert: It’s not just what I’ve taught them but how are they achieving those standards. And always finding a way to assess them fairly but also to…one of our ESLRs – Complex Thinker – how do you assess that, and being able to show higher-order thinking and being able to assess that in a discussion, assess it in an essay, can I do both? But being able to assess that…I mean I won’t know unless I assess. Otherwise, I’m just talking to myself or asking someone to read something.

He was very aware of his shortcomings in the arena of assessment. In response to a question about what is missing from his knowledge base about assessment, he’s very clear that he is missing the technical skills.

The engineering. I think that’s a science and an art. I think I have the art down but not the science. I’m not always aware of every point on the percentile, or whatever

However, once he began to look at the results across the different assessments, his enthusiasm for solving the riddle of a student’s reading ability bubbled to the surface. Like a kid in a candy store, reading the table of results was a joyous process, filled with knowledge about the students and perceptive statements about the intricacies of the interpretation.

Albert: He’s above proficient - 379.
Ted: On the CST?
Albert: Yeah, on the CST. And then on level of contextual silent reading: 6.7, then 10.7 on word. I would expect—you know, something…This is interesting. Oh, here it is. I think the clue to all these…no, it doesn’t work for this one, though—is that the MARSI is showing if they self-tested honestly, I think it’s showing…it indicates some problems in reading if it’s below 60. But this kid, I wouldn’t have expected him to have such a difficult time with this. 6.7. Wow. And then this kid…Oh, this is Linda. Oh yeah. Okay. Linda, she’s also a resource.

As he learned more about the reading construct and the design of the assessments, he was able to make some very incisive comments about the item designs as well as student reading development. For example, during the second interview, we discussed the design of each assessment and what part of the reading construct it was measuring. His comment about nuanced meanings in the vocabulary quizzes -
You’re looking for the words’ meaning and those meanings were nuanced. So I can see the challenge there – that’s looking at nuance – led to a broader conversation about the precision required to derive appropriate meaning from a word in a passage’s context.

As a result of deep examination of the MARSI and its results, he began to recognize that the way a teacher frames a task, a question, or a prompt impacts significantly the kind of data that he gets back. In particular, he became more tuned in to how the students were making sense of the way he was setting up the learning tasks, rather than simply giving a directive and waiting to see the results.

I’ve stopped to listen to the students instead of just giving a direction. When I say summarize, what do I mean, and we can talk about it. And then I realize, oh, that’s not what I mean, they think I mean something different. So we begin talking about what I’m looking for at least… I think what one thing I come out with MARSI, with everything I’ve done with what I’m talking about, is that if I say summarize, it’s not that simple. That I need to scaffold in… here are the steps to summary…Or here’s the kind of summary I mean.

Albert was very enthusiastic about experimenting with his expanded PCK in reading and engaged with new approaches to instruction, recognizing that he still had a lot to learn, particularly in the teaching of vocabulary. His excitement about the “interconnectedness…the word web” rubbed off on his students quickly as did an expanded sense of purpose for vocabulary study.

I asked the students to take two or three words that they struggled with on the vocabulary test and then put them in a word web, which I created…They loved the visual effect, first of all…And they loved the idea that these words have all sorts of… you know. They went to antonyms, synonyms, to other word families. They just… “Wow, okay. So I know one word and I know a lot of words.” And the nuances. And then place different context in which you use them. So connotative distinctions. I realize that vocabulary in that sense -- interconnectedness and context -- is extremely important to teach our students when we’re getting them prepared for the EAP, because that choice diction is really about knowing which word to use. And when you’re writing an in class essay like the EAP, you don’t have a lot of time.

He had a lot of plans for integrating some of the assessments into his regular instruction in the next school year, particularly the MARSI and a much greater focus on vocabulary instruction, particularly the use of word webs and understanding nuance and precision of meaning. In the end, he commented on the challenge of organizing effective instruction. He said, “And one last thing, that education is actually a complex, difficult profession if you do it right.”
Janis. Janis was the one individual who was most tuned in to the students and how the assessment results were impacted by their particular characteristics, such as language ability, background knowledge, etc. For example, when I asked which students would be challenged by the reading fluency tests with all of their little slash marks and lack of punctuation and spacing, she was the only one who noticed that some students with visual processing issues would have a hard time. “Dyslexic students, they would absolutely hate it.”

She also reflected on assessments she had given, trying to understand how students may have processed the test and what it meant.

Janis: I was thinking in my head about a quiz I had given a few weeks ago where I was so surprised that every student got one of the questions wrong and I thought it was ridiculously obvious, but when I thought about it – you know what, that’s actually not a fair question… I do think that reading is still a mystery in a lot of ways and I’m interested in figuring out more about assessment.

Eight years as a teacher, no advanced degrees, a fair amount of professional development in the reading domain, and a mid-range ALI of 74%, she was, in many ways, an ideal candidate for being a subject in this research. She is a good teacher who cares a lot about doing well and serving her students, for example, thinking about how her particular students respond to reading strategy instruction contextually.

Janis: I have taught this stuff explicitly in the past, except I found—and I’ve said this before—like when I teach it to the population here, they’re very antsy, they’re like “Let’s just read.” They don’t like the feeling of going through a process...It’s just they want to skip steps – that’s what it is. Because they feel like they’re smart and they’re ready and they just want to skip steps. I mean, if I did that… It worked when I did one…unit with Shakespeare in it, because that is very challenging to them, and they do want to go through a process and learn a process for it.

As the study progressed, her answers to the interview questions were competent but never really evolved. Initially, she probed a bit into the vocabulary quiz item design, struggling to grasp what the different question designs meant.

Janis: Like, so you’re not just—the question is… it’s not just that you’re testing vocabulary, but you’re testing different ways of getting that vocabulary.

Unfortunately, near the end of the research window, she learned that she was not going to be rehired back into her position (not her choice) so she withdrew from the study. Thus, she did not participate in the second administration of the assessments nor did she want to be interviewed for a final time. She did not take the ALI a second time.
Gene. Another veteran, Gene has a Masters and a fair amount of professional development in reading. Despite that, his initial ALI was the lowest of all of the teachers at 54%. His approach to teaching, and particularly assessment, was very different from anyone else in that he didn’t really believe in testing. Generally, his stance was that testing was not a good use of classroom time, especially compared to performance-based assessment.

*I just don’t give a lot of tests…period. I typically give a lot of projects, assignments, group projects, pair projects, individual responses…If you’re asking 30 students in a class to take one hour, that’s 30 hours of work, what are they going to get out of it? What are they personally going to gain from it? Should they be doing something else? Should they be performing something or creating something that’s more interesting than taking a test?*

Gene’s perspective was testing served bureaucratic reasons for reporting results rather than determining what students know and are able to do.

*I don’t think kids should have to study for an assessment. That’s like going on a crash diet before you go to your class reunion. Does that represent what you know? It represents what you’re forcing out, you know. As opposed to a more natural thing that I should be able to ask them anytime. Cramming and studying facts, I don’t get that. I don’t believe in that.*

He was the first to volunteer to participate in the study but after the initial interview and administration of two of the assessments, his participation waned precipitously. Despite repeated efforts to reengage him, I dropped him from the study.

**Qualitative Data Analysis**

An action research study most assuredly takes place in the real world of school with all of the changes in conditions that are inherent to that environment. The main part of this particular study took place during the Spring semester, when classroom instructional time is already usurped by the preparation and administration of a myriad of standardized assessments including the California State testing (STAR), the Early Assessment Program of the California State University (which impacted the 11th graders) and Advanced Placement testing.

Teachers expressed concern about the time it would take to administer the assessments that were part of this study and expressed frustration about the lack of time to really experiment with instructional strategies that matched up with what they learned about students through the research process. Several felt that the knowledge they gained would have more of an impact at the beginning of the following year, when they were really setting up the instructional context for the whole year.

*Greg: “It’s something…that I think will affect my teaching next year more than it affected my teaching this year…At the beginning of the year, I feel like it’s time for a lot of groundwork, laying out a bunch of stuff. So, I think that using*
the MARSI at the beginning of the year and using it as a teaching tool from the beginning of the year is something...I imagine I see myself incorporating into my teaching. But the changes for this year were more informal.”

The middle school, from which there was an initial promise of up to 5 teacher subjects, was faced with Program Improvement (PI)\(^3\) for the first time. As a result, 3 of the potential subjects withdrew their interest and the other two were challenged to stay focused on the issues we raised through the research.

*Betsy: These big tests that we're held accountable for, versus life skills, critical thinking, and all those other social skill aspect. There are all these elements that go into what happens in the classroom. So, you're juggling...*  
*Ted: So many priorities.*  
*Betsy: Yeah. Prioritizing...I think it's hard.*

Personnel issues complicated the study. In particular, one of the subjects (Janis) was a teacher with a one-year contract. Towards the end of data collection, she was informed that her contract was not extended for another year and she ceased her participation at that point.

Nevertheless, through three different interview sets, two administrations of reading assessments separated by several months, and a brief session of instruction about research findings on the reading construct, the six subjects (4 high school, 2 middle school) generated a lot of useful and important data to inform the research questions posed earlier in this paper. It is to this data that we now turn.

**At the beginning - low Pedagogical Content Knowledge.** The initial interviews showed a clear pattern of low PCK in reading. The three core elements of PCK – pedagogy, subject-area knowledge, and contextual understanding (which includes student background) – were all poorly developed.

The link between a teacher's understanding of a student's background and the depth of their awareness of a student's place on a developmental continuum relevant to a subject-area is particularly important. In other words, the more a teacher knows a student, the deeper the potential PCK. Knowledge of the student's background, his range of abilities, interest and motivation, is a key element in the type of inferences made about the results of an assessment and what to do with that information.

*Albert: As a classroom teacher... you have to be aware of the students, where their strengths are and where their weaknesses are and obviously shape instruction in that direction. So, that awareness - students’ strengths and weaknesses - you also have to connect that awareness to how you're instructing and how you're assessing.*

---

\(^3\) All schools and LEAs that do not make Adequate Yearly Progress are identified for Program Improvement under NCLB. Program Improvement schools and LEAs must implement required program components and interventions.
However, their lack of experience with comparing data on the same construct and really digging in to understand what it means for an individual student was evident in the initial interviews. Even when they knew something about the student’s background (e.g. Special Education, English Learner), they had a very hard time putting all of the data together in a way that made sense and guided them in their response.

*Janis: One thing he's got is severe anxiety and obsessive compulsive disorder... And he cannot do... fine motor skills are extremely hard for him, so this would not... I don't think it all reflects his reading, it (the TOSCRF – silent reading fluency test) reflects his fine motor skills and also just his stress. So... but then as you can see... that's nine out of ten, right (referencing the vocabulary test)? And then this (the MARSI) is fairly high relative to the other students.

But he's struggling so much in school right now that... And I know this from-he's in my advisory and I talk to him all the time-he can't keep up with the pace of anything. His 504 gives him time and a half, but that's nowhere near what he needs to produce the type of work he needs to produce in short periods of time. I really struggle as far as what to do with him to help him.

Watching the teachers look at the reports of the results of the multiple assessments, identifying surprises, inconsistencies, making inferences, etc., I realized that secondary teachers very rarely compare data on a set of assessments that all purportedly measure the same construct. Their domain knowledge of reading is quite weak and most assessments they see are such a conglomeration of skills and knowledge, that teachers are limited to developing a set of impressions about some amorphous, general skill or knowledge area.

Rarely is there a sustained sampling of student skill/knowledge focused on a particular part of the domain over time, i.e. monitoring student development. Secondary teachers are very good at giving assessments, recording grades, and adding them up at the end for report card purposes (which also includes all kinds of data that has nothing to do with proficiency in a particular part of a construct, such as homework completion rates, attendance, participation, etc.), but they have almost no experience with developmental monitoring and analysis.

*Greg: I'm a pretty hard grader on essays but I also give a lot of effort-based work, that really encourages students to build their skills, I think, and also to buoy their grades. So, I think giving different ways for struggling students to build their grades against these kinds of assessments that they're just behind on in a heterogeneous classroom. There's not a lot on different ways of assessing and I don't know if that's something I'm great at, but it is something I'm aware of. Like how do you provide different options for assessment for those who may not be successful at traditional assessments.
Nor do they have access to useful tools such as data collection and analysis systems and accompanying professional development. But it is something that these teachers found useful when they had the opportunity.

_Betsy_: You know, we always talk about multiple measures being important to just not pigeonhole students and make judgments quickly. What makes it hard for teachers is they don't usually see it all like this.

_Greg_: I feel like this data's really valuable when you're sort of looking at it all together, and you can really compare and contrast. Like we're doing sort of across student by student, like one place.

Despite the appreciation of the ability to look at a cross-section of data on a particular skill (reading in this case), their early interpretations were elementary. There was a trust in the test-makers’ definitions of success, such as their evaluation of grade-level results as a key indicator of concern for a particular student, rather than examining the particular elements within the assessment and how it matched up with the student.

_Betsy_: I know that this typically tends to be a little inflated, but I look at it as anything grade level or below is another flag to kind of look deeper.

_Greg_: Students who are reading at extremely high from 11th to the top of the top levels. I mean, a lot of kids who are above grade level… a significant chunk of them, yeah. So patterns wise, that would be the first thing that would jump out at me, certainly the number of really high scoring kids.

For the most part, it's the scores that count in the minds of teachers, not the underlying construct. For example, initially, I didn't tell teachers about how the different questions on the vocabulary test were created - the fact that there were four different types of questions, each with different levels of context built into the question. The way teachers spoke about the results from the 1st administration of vocabulary tests ranged from blaming themselves for not teaching the words properly to surprise that the students had forgotten so much despite multiple strategies to teach the words. Nobody had really looked at the construct of each question.

_Sandy_: When I looked at these (the vocabulary tests), there were definitely some surprises. Some kids that were much higher than I was expecting and kids that were much lower than I was expecting. And I don't know what the absolute numbers mean - it was all for relative comparisons - but I was really surprised at times. And so I kept trying to figure out what exactly this was telling me. And I kept not being sure if I was making the accurate inferences based upon this data.

A particularly interesting challenge in reference to student background was the nature of the student population in Albany. It is generally a high achieving district with a lot of good readers. The higher the reading skill level, the more difficult to evaluate a
students' reading because parsing out the elemental skills is more challenging. There is so much integration of the key skills at a higher level. So in Albany with so many strong readers (a grade level reader in the AUSD is considered low end), not only is it more challenging to figure out what a student needs but there is less motivation to do so. A student who is scoring in the proficient or higher range isn’t generally a concern. Teachers don’t need to drill down and try to understand what their reading deficits are.

_Betsy:_ We talk about what separates an honors kid from a regular kid, and that even if we never taught these skills, the kids who...are really the highest skilled kids. They use those strategies whether they’re fully conscious of it or not.

The Assessment Literacy Inventory (ALI) provided an important look into how teachers thought about assessment. The ALI was administered twice – once at the beginning of each teacher’s involvement in the study, and again at the end. While there was only a slight gain among the teachers who took it both times (see Table C below), it served to set the stage for deeper exploration and self-reflection about assessments that carried through the study.

The interviews following the first administration of the ALI illustrated that teachers were challenged (and a bit frustrated) by the academic language of the inventory.

_Greg:_ I learned for one thing there is definitely some pedagogical language with which I am unfamiliar and my reaction to that was really frustrating because I felt that someone fresh out of their teacher education program would be able to answer the questions more correctly and better than a veteran of 20 years because of better understanding of some of that language.

_Albert:_ It was actually difficult because thinking about the language you use as an educator with testing and assessments is rarely shared enough among teachers and we don’t use the language a lot. And you don’t use it with students. So using it, reading it seeing it, thinking about it. It was, I thought, very challenging.

Yet, it caused them to wonder how it reflected their own knowledge and familiarity with assessments and their pedagogy in general.

_Betsy:_ First of all it was surprising. It was interesting to see the teaching experience distilled down to a multiple-choice format.

_Janis:_ I had to kind of review the concepts and it made me think to myself, for example norm-referenced vs. criterion, holistic rubric vs. whatever the other type is. So, a lot of the terminology, I hadn’t actually accessed since my credential program and even then, it wasn’t as emphasized as some other things. So, it made me think...once I reviewed the concept it made me think, do I do this anyway? Do I
formulate valid and reliable assessments’ anyway without having that terminology attached to it? Or does it help me to review those concepts.

**Struggling to make sense.** Despite the relative lack of knowledge about PCK (including assessments), there was strong evidence that the teachers understood that these are complex and important issues for them professionally. Their motivation to learn about the different aspects of the reading domain, its assessment and instructional strategies came from a deep respect for the students and their time and needs. There was also a conceptual grasp of the idea that the lines between assessment and instruction are often blurred, as in how a teacher assesses reflects how he teaches.

Albert: You have to be able to create that instruction as well as that assessment, that engages the student to do something but also measures the students ability to do it…That kind of engineering is probably more challenging than we give it credit for. Creating a rubric, for instance, sitting down and thinking about what a student should know so they can think ahead and say here's what I'm learning, here's what I have to do.

The interview questions caused the teachers to think more deeply about the relevance of a given assessment to a subject-area construct (i.e. what assessments are actually assessing and trying to be sure they are an accurate reflection of the skills or knowledge targeted). Unfortunately, this is not as common a process as one might think. As stated previously, teachers in the modern secondary public school environment rarely have the time or tools to engage in such deep thinking about what a particular assessment measures, much less what it means for a specific student. But the opportunity to analyze an assessment tool revealed some perceptive thinking.

Sandy: I was thinking in my head about a quiz I had given a few weeks ago where I was so surprised that every student got one of the questions wrong and I thought it was ridiculously obvious, but when I thought about it - you know what, that actually was not a fair question.

Albert: (Referring to the silent reading fluency measure) It tells me that they don't see patterns in language and that bothers me. That tells me they haven't got the basics down yet. Like I'm going to compare it to chess. When you first start chess, if you don't think about every move, step-by-step, it's likely you'll lose against someone who is good at it. But once you get those moves down, you start to see patterns developing. I think it's the same with language. You have these basics and you understand how word order works, you anticipate that. You understand parts of speech to a certain degree and you understand meaning as well. You start to see meaning in that.
For one teacher, her thinking about the strength of an association between an assessment and the subject-area construct was represented by teacher confidence about the subject and its pedagogy.

*Betsy:* I think what makes a teacher good at designing assessments is a real confidence in what the underlying goals are for the lesson, for the curriculum, and a good knowledge of their students' abilities and level of development.

*Ted:* You said the word confidence. What do you mean by that?

*Betsy:* Well… I don't think enough of us really feel confident about the goal that they're trying to teach. I think it's very blurry. Let's take a piece of literature. We know what we want students to come out with generally. There are a lot of things that go into teaching literature. But how confident do you feel about prioritizing? What is the most important thing... what skill, what knowledge, what facts... How do you prioritize what really is the goal of teaching that? I think we don't feel confident. Maybe confidence is not the right word.

Another teacher made a comment that reflected his understanding that the elements considered in the design of an assessment item were critical to its effectiveness as a measure of a particular skill. How a question is set up, what its context is, the test directions, and its timing in a unit are all factors that affect the inferences that might be made.

*Greg:* Clarity helps a lot.

*Ted:* Clarity of what?

*Greg:* Of expectations. I mean when I think of assessments, I generally think of essays but it's obviously much more than that. But the instructions, the prompt, expectations being really clear.

**Improvement in analytical depth and sophistication.** There is a difference between theoretical, pedagogically rich discussions of assessments and those just responding to data. The first is reflective of a more developed PCK, the second less. A pattern emerged as the study progressed that higher PCK lends to more sophisticated analysis of the implications of assessment results for inference and instruction.

**Cognitive and Metacognitive Strategies.** For example, the construct of reading includes a major element of cognitive and metacognitive strategies. As teachers examined their own strategies, and analyzed student results on the MARSI, a deeper understanding of both the construct and student development emerged.

The self-reflection of the MARSI caused a bit of self-reflection among the teachers and challenged the assumptions they make about how students read, illustrating the overwhelming complexity of the reading construct. Betsy, in particular, came to a realization that teachers may know a lot of the metacognitive strategies but...
they often avoid the breadth and depth of analysis of student reading behavior that they represent.

*Betsy:* (As I took the MARSI) I thought of my students. I thought about how I teach. I thought about what I see them do and not do. It was actually a little bit concerning to me because there’s so much that they don’t do, or that I don’t see them doing, or it’s not clear that they use strategies like this. And that we assume so much.

*Ted:* We assume it, why? Why do you think we assume?

*Betsy:* I think we assume it because we don’t want to break it down. I think a lot of us don’t really think about it in this level of detail. So it’s almost as though… ‘Well, I don’t know it, so I don’t want to think about it, because it’s a lot to think about.’ It’s a lot. These are 30 strategies.

Betsy’s comment illustrates a key point in my analysis. Teachers and researchers (and the public) may not want to deal with the complexity of learning and assessment because it has too many moving parts. It is too complicated, too hard to make sense of. It is much easier to simplify and boil everything down to a simple rating or grade or score. But reducing a complex skill such as reading and evaluating someone’s reading skill according to one number such as a grade level is a disservice to the student, teacher, and ultimately the society that is interested in children learning to read.

This metacognitive strategies assessment, more so than any of the other reading tests that teachers took, stimulated the richest discussion about instruction. They were inspired to think a lot about how they could use the assessment in their own classrooms and made plans to teach around the results right away.

*Albert:* I have such a range with sophomores and there are students absolutely lost in the first paragraph and it befuddles me. And I really think a strategy, a tool like this could take them out of that and make them feel empowered. Make them feel like at least they can get into the text.

One of the teachers, after administering the MARSI to students twice and more deeply understanding its relevance to the reading construct, saw the self-reflective nature of the assessment as a particularly useful tool to build upon and increase student engagement and independence as readers

*Greg:* I think that in the future, doing the MARSI, administering the MARSI at the beginning of the year and actually talking more explicitly through all those things and what they mean. Allowing kids to even change some of their answers once they have a better understanding of what those things are.

**Vocabulary.** Another key element in the reading domain is vocabulary, which is, in itself, a remarkably complex structure. The influence of context on the meaning of a word, the background knowledge needed by the reader to grasp nuanced meanings,
and the precision of interpretation necessary to meet the readers' purposes are just some of the many variables that readers and teachers of reading must navigate. How readers build vocabulary and how it is taught has been the subject of a great many books and articles. The assessment of it, given its complexities, poses extraordinary difficulties. The multiple choice vocabulary tests that I constructed for the teachers were drawn from four different item designs, each representing differing levels of context present in the question. The fact is that none of the teachers noticed the variation in the types of questions. This had a major impact on the type and quality of their initial inferences about student reading ability. With the first administration of the vocabulary tests, teachers simply looked at the scores, and as described previously, tended to use it as a somewhat confusing, general representation of reading ability.

Janis. These scores are lower than I would want them to be...There's a varying level of difficulty. But we did study these words; they wrote sentences with them. I think it was part of one of the quizzes. So I wonder about retention of words and if there are things you can do to...you know, how do you retain vocabulary?

While the teachers struggled with understanding the implications of the results from the first administration – for example, the teacher in the example below points to background knowledge and its role in understanding context - their perspective was skewed by the raw score. They did not question the difference among the test items themselves but were beginning to make connections between a growing understanding of the reading domain and assessment results, thinking about what kind of instructional strategies make sense.

Albert: So if they have to look at a word in context to understand its meaning - to derive meaning of the entire piece or the entire sentence or the passage - that's where they struggle. That's what it tells me. I mean, even though we had some 8's and 9's and a 10. Well, okay. Was that the only 10?

Ted: Yeah, there weren't a lot of 10's.

Albert: So, I think that's the background knowledge. So I'm saying two things: (as a result of the MARSI, their) strategic reading strategies are beefed up, stronger, more muscular, and at the same time, I think there's reading strategies they could use. And then vocabulary or background knowledge...I think that can be stronger. But the background knowledge...they could go together, but I don't know...I think sometimes you have to focus on them in isolation and then bring them together.

Between the interview following the first administration and the second administration of the vocabulary test, teachers received professional development on the reading domain (see section in Methods above). They also read an article that described the four different vocabulary item designs for the vocabulary test. Their
expanded subject-area knowledge was dramatically reflected in the third and final interview, examining the results of the second administration.

For most of the teachers, they had a new motivation to go deeper in vocabulary instruction with their students and approach it in a very different way.

*Sandy:* For me, having this experience makes me want to spend more time with individual words and really sort of focus my teaching more to really build around key vocabulary words and concepts as opposed to thinking about it as like a list or an outline of five or six key ideas. So building from within as opposed to moving down a list.

*Betsy:* That article you gave me that broke down the vocabulary levels, ways of understanding the words was really new to me. Let’s put it this way: it changed my method of teaching.

*Albert:* This has been actually a… I would say a major influence in terms of thinking about words and teaching words… teaching vocabulary. But how to teach it and what situations and what purposes. And why are students using vocabulary and making sure students understand that.

As Betsy said, preexisting teaching methods were challenged and the teachers explored ways to teach students the non-linear quality of word meaning. Their newfound knowledge of how the different question types captured the nuances of word meaning led several to experiment with word webs as a way to help students see how context impacts not just word choice but subtle differences in meaning.

*Albert:* I took them and I asked the students to take two or three words that they struggled with on the vocabulary test and then put them in a word web, which I created.

*Ted:* And how did they find that? How did you find that?

*Albert:* Really, really enjoyed it. They loved the visual effect, first of all… They loved the idea that these words have all sorts of… you know. They went to antonyms, synonyms, to other word families. They just… “Wow, okay. So I know one word and I know a lot of words.” And the nuances, connotative distinctions. I realize that vocabulary in that sense -- interconnectedness and context -- is extremely important to teach our students when we’re getting them prepared for the EAP, because that choice diction is really about knowing which word to use.

**Students making meaning from text: How precise does it need to be?** As the teachers developed in their sophistication with reading strategies and vocabulary, the conversations took an intriguing turn, into an area that may have interesting implications for instruction and assessment. A question arose about how precise does
a readers’ understanding of a particular word have to be to derive sufficient meaning from a text?

Reading is about inference derived from text. A reader uses a variety of strategies to put together their knowledge of grammar and syntax, contextual understanding, and background knowledge to derive meaning from the words on the page.

On a typical reading test, there’s a requirement for precision. Usually, there is one correct answer and in order to get that one, precise answer, you have to read carefully. Whereas if you’re reading a paragraph in a novel, most of the time, if you have a general sense of a word’s meaning, then you’ve got the gist of it and that impressionistic understanding is good enough in the context of the whole paragraph and the story arc.

The purpose for reading affects the degree of precision of meaning required. For example, the word “grudge” was the focus of one of the vocabulary tests. When a reader sees the word in a sentence as part of a paragraph in a story, it has characters and context surrounding it that’s related to plot, setting, and other story elements. In this case, it’s helpful if the reader is familiar with the word, but he does not need to know exactly what the word means to be able to derive the meaning that’s appropriate for moving forward with the story.

On the other hand, on a vocabulary quiz, the precision of meaning required to answer a question correctly can be quite different. But even on a quiz, the design of the question affects the precision required. The two multiple-choice examples below illustrate different levels of precision required to get the right answer\(^4\). The first requires a fair amount of familiarity with the word because there is no context. However, there is no need for real precision because the incorrect choices aren’t even close to the meaning of the word. So, if a student’s word knowledge is in the ballpark, they will get the answer correct.

1) **grudge**
   a) good will
   b) oily substance
   c) bitterness
   d) obligation

   The second example is contextual so it appears that a reader’s understanding can be more impressionistic because there are clues in the sentence. However, a close reading shows that three of the four answers could be inferred to be the correct answer. It’s deceptive because the context clues are fairly nuanced so the meaning is quite precise and only one answer (resentment) is given credit.

2) It was hard for the student to hold on to a **grudge** when the teacher was so fair.
   a) sadness
   b) envy
   c) resentment
   d) favorite

---

\(^4\) These questions are copied from the actual vocabulary quizzes created for this teacher as part of this study and use item design strategies described in the methods section.
The idea of a spectrum of meaning inference from precise to impressionistic stimulated a lot of discussion among the teachers. Greg was concerned that students were tripped up by a test that had variations in the precision of meaning required to get the right answer.

Greg: It's almost like that precision inhibits them a little bit. Like they know exactly what I say the word means and then they look on the test and they don't see that exact word. And some kids work it out and some kids get flustered. So in some weird way, that precision almost hurts them in that kind of test where there are so many definitions for some of those words.

There was a fair amount of thought about what kind of instructional approaches might teach students to manage this spectrum of precision, i.e. knowing which level of precision is required for a particular text or purpose. Betsy, while examining the results from one particular student began to generalize about appropriate instruction for word complexity.

Ted: So she doesn't use a lot of strategies.
Betsy: Doesn't use a lot of strategies.
Ted: Perhaps it's the more complicated words that might be part of the issue.
Betsy: But then again, she does. She's a person, if I teach her a strategy, she will go home and study it. And she will use it. So what I'm thinking that these results are showing me is that this needs to be more emphasized and there needs to be more time spent on...when the complexity of the word is higher, I need to spend more time teaching and practicing it in a variety of contexts.

Albert, on the other hand, was excited by the implications for instruction and students growing as readers, drawing a connection to science reading.

Albert: I loved your scale from impressionistic to very precise. If we can move the students in that direction, I think it makes them powerful readers and writers... And I think that's why a lot of them love the sciences - the advanced classes - because it gets very precise. And it's definite. Those definitions are clear cut or much clearer.

As stated above, the degree of precision of meaning required depends upon the purpose of the reading. A neurologist reading a lab report from an MRI requires a very precise understanding of every word in the report. Lives depend on it. Whereas, a 7th grade science student scanning an article from the newspaper to meet the requirements of a current events homework assignment worth 10 points (out of 300 over the course of a semester) has a very different requirement for understanding the meaning of all of the words in the article. While these examples are obviously at opposite ends of the spectrum, they do illustrate the challenges in designing valid assessments of reading.
Restructuring Conceptual Frameworks. In the third and final round of interviews (which occurred after the second round of assessments), there was increasing evidence of greater complexity of thought and deeper understanding of the issues. Teachers had time to reflect on first round of assessments, take them themselves, learn more about the reading construct, and then administer the assessments again. A new conceptual framework emerged from their growing understanding about the reading domain and how different assessments fit the various elements in the reading construct. This resulted in more creative thinking about what aligned instructional strategies might look like and how to prepare students for the assessments better. While there was not really enough time to see significant instructional changes for the current school year, there was a lot of talk about changes planned for the following year.

Some of the teachers were able to reflect back and summarize what they had learned through their participation in the research process. Their comments reflected an expansion of their conceptual framework for reading and how different elements of the domain and relevant instruction and assessment fit in.

Sandy: Overall, I think I'm just more aware of trying to assess all of the different facets of reading and trying to do that in the beginning and at the end. Which I think I knew before, but I hadn't seen it broken down into all four different kinds of questions that are available, and all of the skills that are really laid out on the MARSI. So I think just sort of having access to… These materials and the different kinds of questions is really… just sort of helps me specify exactly what I'm doing, which I think is really helpful.

Betsy: Definitely I've changed. I have changed in my awareness of specific strategies and specific areas that comprise the whole comprehension process. It really made me aware of it.

They were able to talk about individual students and the impact that the different assessments and subsequent instruction had on their orientation to learning, how the students themselves were becoming more conscious of what was involved with reading.

Albert: This student I noticed specifically because I'd also been working with him. And I've noticed he took to some of the ideas and techniques I've been focusing on.

Ted: What do you mean by that? He took to…

Albert: Well, he seems to take them seriously. When we

---

5 There is a similar spectrum for assessments. Teachers are trying to derive meaning from an assessment that helps them understand the students' knowledge, achievement, or skill and then make decisions based on that. Different learning goals will require different degrees of inference precision.
talked about summary or when we talked about finding evidence and claim, he was writing this down; he was taking it seriously. And then for his vocab word web, he did a really good job with that. He even expressed to me he felt this was good, this was good practice for it, and even said that he wished he could do more of this.

Their thinking about instructional approaches for the new school year reflected bits of experimentation with different approaches during the current year as a result of this investigation.

_Betsy_: I had been trying to do a lot of teach… not a lot… I have been, in the back of my mind, I was conscious of all of those metacognitive strategies around reading…I did not systematically look at ways to - besides just telling and talking to the kids about it - I didn't really do much around teaching or instructing. That part I don't think I really did. And I'm trying to think on how to do that for next year because I think I'm going to give that MARSI to every student. All my English classes at the beginning of the year and at the end of the year.

In terms of reading fluency, both Greg and Betsy struggled with both the interpretation of results and figuring out instructional approaches that made sense.

_Ted_: then there’s the fluency. How much can they read through and be comfortable to be able to try to capture that essence. It's really, it's an interesting challenge.

_Betsy_: Yeah, I think what I'm going to do is scaffold it with… Beginning of the year… small paragraphs, and then sections, then pages, then chapters kind of thing. Well, I don't think we'll ever get to chapters. But just trying to work up from a small… maybe even a sentence, couple of sentences and then moving up. Then I'm really going to have to figure out the vocabulary. Because the vocabulary, to do it right, just slows you down. It's that pacing. It's really hard.

_Greg_: I feel like we were talking about actually picking up the actual fluency, the pace of reading. And I don't really know how to instruct that at all. As an English teacher, I'm like well, it's not math or science, you're supposed to read all the words, you're supposed to try to enjoy it a little bit. And you said something that's to the effect of ‘oh yeah these kids need some strategies for getting through all the books that they're going to face in college’. And of course that makes perfect sense, but I don't know how to pick up the - not skimming and stuff like that - but actually picking up the actual fluency of reading, every word reading, the speed of reading, every word reading… I don't really know how to do that.
Sandy had some ideas about teaching vocabulary but was flustered by the extraordinary complexity of the task.

Sandy: I think there's something inherently really, really hard about analyzing something that complex (vocabulary knowledge) through a multiple choice test...I think ideally, if I were really to assess knowledge of a word, I would want them to create something like a word web or something so they could - they were generating the different meanings and to show their understanding of the word that way.

A particularly interesting side-effect of this work was how the students became more attentive to the facets of the skills being measured as the teachers become more familiar with the assessments and then talked about them with the students. The MARSI itself is a self-reflective assessment which, if discussed with the class, would naturally lead to greater familiarity with reading strategies. But what also changed is that the teachers began to use a more informed and sophisticated language about reading and the students followed suit. Albert pointed out that for some, it gave them a way to manage their own struggles with reading.

Albert: I have such a range with sophomores and there are students absolutely lost in the first paragraph and it befuddles me. And I really think a strategy, a tool like this could take them out of that and make them feel empowered. Make them feel like at least they can get into the text.

Sandy was excited about the use of the MARSI in particular as a springboard for conversation but was suspicious about its use as a true measure of student reading skill.

Sandy: The MARSI being reflective is... I think trying to assess a kid's awareness of their reading strategies, which is a little bit complicated because whether you're aware of them or not, you may not have them reading strategies wise. And just because you're aware of them doesn't mean you're actually employing them effectively. It's a really good thing and a great springboard for conversation, but to me, like a slightly less accurate judgment of being able to take that step and apply it to saying a kid is a great reader or not.

Teachers don't often examine and compare the results of an array of standardized assessments of the same skill or knowledge. When they do, it's an opportunity to try to make sense of apparent inconsistencies or contradictions that may become evident. This deeper knowledge of a student strengthens their PCK and subsequent capacity to intervene instructionally. It also helps them realize that the assessment process is quite complex with many facets that they don't normally address.

Albert: I do look at the numbers for 12 graders who took CAHSEE, then STAR testing. So, I look at those numbers, I have a sense of where they are but sometimes I feel like it's just telling me grade level stuff. It doesn't tell me enough. It
seems like I get a lot more. This doesn’t represent enough of what I do in my own classroom. I feel like I get a lot more with my own observation and techniques. Sandy: It’s just so interesting to see how… definitely for a few of them. The contrast between how I,… what I have seen from their sort of metacognitive experience and the self-reporting that happens. And also, some of them really stand out as testing much, much better than they perform in class.

Their struggle to make sense of the contradictory results creates an opening for professional growth but the lack of time for that in the public school teacher’s day and a general lack of cultural norms supporting such growth means an opportunity is wasted.

**Growth in PCK, increased motivation to expand instructional strategies.** A teacher who has a well-developed PCK with all of the pieces in place, may not have a lot of instructional strategies to choose from to meet student group or individual needs. A relatively inexperienced teacher faces this no matter how intelligent they are.

*Betsy: My gaps continue to be vocabulary. Teaching vocabulary effectively and assessing that aspect of reading comprehension… assessing it properly. I think probably the assessing piece, really.*

*Ted: Specifically around vocabulary or in general?*

*II: No, in general… in reading comprehension. And then the follow up based on the assessment results. I feel that I’m not assessing quite the way I should. And then I’m not taking those results and re-teaching or changing my approach based on those results. Like the chicken and egg, I’ve got to do kind of a better job of assessing before you have something to go back and change your method.*

Teachers who have worked in the same fashion for years have a very difficult time figuring out what to do with their newfound conceptual framework of the reading construct. The transition from a pattern of little self-reflection and little guidance about how to navigate a particular subject-area lends to a set of instructional strategies that have been picked up haphazardly through a variety of means, from “this is how I was taught” to “I have a friend who gave me this lesson idea.” Rarely is there the framework of research-based professional development that is sustained over months with many hands-on aspects led by the Principal of your school. Such an environment definitely stirs up and exposes the lack of viable options in a teacher’s instructional strategies quiver. The excerpt below is an example of one of the teachers confronting the gaps in his own quiver.

**Greg: The way that you deal with multiple choice is not… I mean, I told you I’ve done some matching and I’ve done some fill in the blank, and I also do some straight “here are the words”. And sometimes I’m like write them in a… use the word in a clear sentence. Whereas I’ve always had this kind of hodgepodge of ways that I go about trying to mix it up so...**
that kids who just memorized the words have a chance to show that at least they prepared...But I never really understood how to do a multiple choice test. It always just seemed like if I made one it would just end up so obvious what the answers are. So... at least trying to construct my own vocabulary test that way is definitely one thing I would do.

Synthesis of qualitative analysis

At the beginning of the data collection timeframe (i.e. the first interview), teacher knowledge of the reading domain was weak, assessments they traditionally used to measure a student’s reading ability failed to distinguish between discrete elements of the reading process, and tools that were available to them for data collection and analysis suffered from challenges of data input and their poor knowledge and experience with interpretation of that data. Thus, teachers were left with an impressionistic view of individual student reading development and were forced to rely on a test-makers single-digit representation of a child’s reading ability (e.g. grade level).

Initially, there was a tendency to engage in simplistic self blame when scores of assessments were not high; as opposed to trying to understand what each assessment was measuring and examine each student and his progress from a variety of skill perspectives. Complicating matters, teachers don’t generally have the time or resources to develop experience and skill with the intricacies of the reading construct, or what Sandy called the “enigma of the reading experience”. This includes not just the different elements in the reading domain but also what different assessments measure and what specific results mean much less time to plan for appropriate and differentiated instruction to meet student need.

As the study progressed and teachers took various assessments themselves, had a chance to reflect on those results, administered some of the assessments a couple of times and received professional development, they expanded their Pedagogical Content Knowledge and assessment literacy. The act of reflection powerful especially around the MARSI with its “30 concrete tasks” as was focus on vocabulary especially after teachers began to understand the design of the test items.

The exploration of the need for precision in making meaning from text that emanated from the vocabulary assessments and subsequent conversations was particularly useful for helping teachers see the depth of complexity in the reading domain and the role of assessment design in eliciting useful information about an individual students’ reading ability

A new confidence emerged from their increased knowledge and sophistication. They became more motivated to meet the challenge of designing new assessment items and figuring out from the evidence what a student is able to do and what is needed to progress.

They began to challenge preexisting teaching strategies. A growing knowledge led to motivation to change practice and experiment with different approaches and strategies. But their evolving grasp of the complexity of their task, if done well, presented a danger of being overwhelmed by their lack of knowledge about instructional
strategies for reading. As stated previously, a focus on developing instructional strategies was not within the scope or time limitations of this study.

**Quantitative Data Analysis**

A clear pattern emerged from this research that increasing experience and close examination of assessment data from a variety of reading assessments, plus professional development on reading and assessment would lead to greater sophistication with the interpretation of those assessments and better instructional decisions. While the qualitative analysis of the coded interview excerpts over the course of the interviews lends credence to this observation (as described above), it was appropriate to conduct a statistical analysis of available data to determine whether there is quantitative evidence as well.

Given the small sample size, particularly among those teachers who participated in all aspects of the study (four of six cases), any statistical result would likely be suspect (particularly any measures of statistical significance) but there are still some interesting results that are worth mentioning.

**Assessment Literacy Inventory.** The most interesting available data to examine was the results of the Assessment Literacy Inventory (ALI) and the changes in the rubric scores over time. The ALI was administered at the beginning of each teacher’s involvement with the study and then again at the end. Hypothetically, as a teacher became more experienced and savvy about assessments in general, their ALI should increase.

In fact, there was a fairly large positive change in the results of the ALI from the first to the second administration. As is illustrated in Table C, the average score of the respondents increased from 73% correct to 77% correct.

The correlation between the two administrations of the test was very high and statistically significant ($r=.97$, $p < .05$). It is important to note that the teachers did not receive the results from the first administration of the ALI nor did they have a copy of the test until after its second administration. Therefore, it is reasonable to conclude that the increase in scores was related to external factors, and not just to chance or familiarity with the test and its result.

<table>
<thead>
<tr>
<th>Table C – Assessment Literacy Inventory Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean increase in ALI scores from 1st to 2nd administration = 5.5%</td>
</tr>
<tr>
<td>Pearson Correlation $r = .97$</td>
</tr>
</tbody>
</table>

**Rubric Scores.** The rubric scores (Pedagogical Content Knowledge and Instructional Decisions; Assessment Literacy and Inference Quality) were applied by the researcher to excerpts from the three interviews. Again, hypothetically, scores from the first interview should be lower than those from subsequent interviews as the teachers grew in their knowledge and sophistication about Pedagogical Content Knowledge, assessments and instruction around reading.

The results did not bear this hypothesis out for PCK and Instructional Decisions (PCK & ID) but it did for Assessment Literacy and Inference Quality (AL & IQ) - see Table D. The mean rubric scores for PCK & ID actually declined slightly between the 1st
and 3rd interviews (-.01 points on a 4 point rubric scale) while the AL & IQ scores increased fairly substantially (+.26 points on the scale).

<table>
<thead>
<tr>
<th>Table D – Rubric Scores</th>
<th>Pedagogical Content Knowledge and Instructional Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview</td>
<td>Mean Score</td>
</tr>
<tr>
<td>#1</td>
<td>2.2</td>
</tr>
<tr>
<td>#2</td>
<td>1.95</td>
</tr>
<tr>
<td>#3</td>
<td>2.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment Literacy and Inference Quality</th>
<th>Interview</th>
<th>Mean Score</th>
<th>Standard Deviation</th>
<th># of Excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>2.2</td>
<td>.51</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>2.15</td>
<td>.59</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>2.46</td>
<td>.77</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

A possible explanation for this relatively flat result in PCK & ID may have something to do with the nature of the excerpts that were scored from the different interviews. The questions from the first interview focused on the teacher’s thinking that resulted from their taking different assessments, including the ALI, the MARSI, and both silent reading fluency tests. This was very early in the process, no student assessments had been administered, and there had been no professional development about reading and instruction. Thus, their Pedagogical Content Knowledge about reading was limited. On the other hand, their answers to questions tended to show a higher degree of sophistication about assessments possibly because they were self-reflecting and knew a lot about themselves as learners and could figure out more about what the assessment results meant.

A very interesting result is the steep drop in scores that occurred from the 1st to the 2nd set of interviews with a loss of .25 points on the PCK & ID scale and .05 points on the AL & IQ scale. The second interview was all about interpreting a complex set of student assessment results for the first time. Now they were trying to understand what was going on with 50 or 60 students, thinking about their different facets of reading skill but also their background, orientation to learning, language challenges, etc. It was a much greater analytical challenge and appears to have caused them to retreat to more elementary thinking about the students and their development as readers.

A complete reversal occurred between the 2nd and 3rd interviews with an increase of .24 points on the PCK & ID scale and .31 points on the AL & IQ scale. The third interviews asked the teachers, once again, to interpret a new, similar set of student assessment results but now they benefitted from professional development about the reading domain and assessments, plus the experience resulting from the first look at the assessments.

I became further intrigued about the correlation between the growth in the teachers’ ALI scores and the increase in rubric scores, particularly between the 2nd and 3rd interviews. Such a correlation would be a useful confirmation of the research results. In fact, there was a fairly high correlation between the ALI results and the AL & IQ data, not surprising given the focus of both measures on Assessment Literacy,
although the result was not statistically significant ($r=.72$, $p>.05$) – See Table E. The correlation between the ALI and the PCK & ID data turns out to be fairly weak ($r=.17$, $p>.05$).

<table>
<thead>
<tr>
<th>ALI Change to:</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL &amp; IQ Change</td>
<td>.72</td>
</tr>
<tr>
<td>PCK &amp; ID Change</td>
<td>.17</td>
</tr>
</tbody>
</table>

These results would lend support to the idea that teachers who take the time to deeply examine a complex set of student data related to a subject-area domain increase their sophistication and expertise in their ability to make appropriate inferences and instructional decisions, particularly when they receive professional development in topics related to that domain.
Chapter 5: Conclusion and Implications

I started the data collection process with two major questions in mind, one that was focused on inferences and one focused on instructional decisions (see Chapter 1) and learned how intertwined the answers to these two questions are.

Secondary teachers start with little experience with assessing a particular construct from a variety of angles with developmental proficiency in mind. Their perspective tends to be somewhat simplistic, without much expertise in the domain, thus they don’t have much of a frame of reference to be able to think about what kind of assessment gives particular information about some facet of a domain. There is little time to reflect on these issues much less the training or the knowledge about what to even consider.

It is clear that as teachers reflected and challenged their assumptions, they began to reconstruct their conceptual frameworks about teaching reading, i.e. they learned more about the construct, experienced different assessments, examined what they were actually measuring, and integrated results of those assessments with the results of other assessments and specific knowledge of students, their inferences about where students were developmentally in the construct became more complex and sophisticated and of a higher quality.

The PCK for the domain of reading is extraordinarily complex and poorly developed by most teachers at the secondary level. The validity of assessments used to measure learner development is dependent upon not only the context in which they are administered (i.e. background knowledge, learning environment, purpose), but the PCK of the user of the assessment (the teacher). For example, a learner may have quite developed strategies for determining meaning from a text but lack the background knowledge to comprehend complex concepts.

Vocabulary development plays a major role in the reading construct, yet clarity about its construct is elusive. Words typically have many nuanced meanings so assessment and instruction is bedeviled by a spectrum of required precision of definition. For example, typical reading of a novel’s paragraph often requires a somewhat impressionistic definition whereas a definition of a term in a legal or medical context likely requires much greater precision in the definition.

The greater the PCK of the teacher, the more fluid they are in their ability to navigate these complexities, balance the often contradictory data streams, and identify instructional strategies that serve both individual students and the group, despite the relentless time and management pressures of the typical teaching work load.

In summary, teachers generally operate at a fairly low level of PCK, thus are limited in their roles as educators to that of lower level instructional guides at best - technocrats implementing what someone else designs. However, the higher the PCK, the richer and more complex their knowledge of the student’s development and the more precise and appropriate the instructional decisions.

The findings from this study have a range of implications in a variety of educational contexts, including accountability, curriculum design, teacher pre-service training, in-service professional development, data warehousing software, and standards and accountability systems.
The “high-capacity” schools with strong internal accountability systems described by Elmore (2004) include teachers with strong pedagogical and assessment skills. Wilson’s “Community of Judgment” (2004) relies on teachers and their ability to make sound judgments as the center of an effective accountability system.

In the current accountability systems, limitations imposed by standardized tests force limitations on teacher development and therefore, student development. If we want an accountability system that truly encourages sophisticated teaching and learning, then more power has to be put in teacher’s hands but with appropriate training and time to develop.

The kind of investigation that teachers engaged in through this research and the reflection on the results of the student assessments, in coordination with basic professional development specific to the reading domain, had important effects on teacher thinking and instructional plans. The participants were energized to try new approaches, their inferences about student learning conveyed a much deeper and richer knowledge of the reading domain and the backgrounds of their students, and their instructional decisions reflected a broader understanding of the developmental construct of reading.

This is what the teachers in a “Community of Judgment” look like. This is what they do. They reflect, they are confused and they dig their way out, they explore, they reinvent, they are engaged. It’s what professionalism is all about.
Appendix 1: Interview Questions

**Interview #1 – ALI/MARSI/TOSCRF Results**

- General Questions
  - How would you describe your general attitude and thinking about assessment?
  - Why did you volunteer to participate in this research?
- Assessment Literacy Inventory (ALI)
  - What was it like taking the ALI? What did you learn about your own thinking about assessments?
  - How important a part of the teacher’s job do you think assessment of student learning is?
  - Do you think it represents a valid assessment of your own assessment development skills? Why, why not? What information is missing?
  - What do you think makes a teacher particularly good at designing assessments?
  - Do you think you are good at it, based on your answer to the previous question?
- Metacognitive Assessment of Reading Skills Inventory (MARSI)
  - What did you realize about yourself as a reader as you were taking the MARSI?
  - Which areas did you rate yourself the highest? Which the lowest? What did you think about as you were rating yourself, i.e. what information did you draw from to make the rating?
  - Did you think about your students as you were taking the MARSI? What were you thinking about? What do you think it tells you about their reading skill?
  - What did the MARSI make you think about instruction? Any sense of the strategies that might follow the results? Do you think you might make such strategies a priority at this time? If yes, why? If no, why not?
- Test of Silent Contextual/Word Reading Fluency (TOSCRF/TOSWRF)
  - Describe what you were thinking about as you were taking these reading tests, about your own reading skill as well as your students.
  - What do you think each of the tests measures, i.e. what reading skills?
  - How important is it to your instruction for you to know the kind of detail you just described?

**Interview #2 – Results from 1st Student Administration**

- Describe briefly each of the tests that were administered and what skills or knowledge you think each is assessing.
- Thinking about the results of all of the tests together, was there anything in particular that surprised you or grabbed your attention?
• Looking at individual student results, did you notice any apparent inconsistencies among the results of the different tests? If so, do you have any thoughts about what may have caused the inconsistencies?
• What do these results suggest to you about how you should be teaching reading to your students as a whole group? As individuals?
• Do you have thoughts about what your next steps will be with your students?

Interview #3 – Results and Instructional Decision-Making from 2\textsuperscript{nd} Administration

• Since we examined the results of the 1\textsuperscript{st} round of reading assessments together, have you actively taught reading to your students?
  ○ If so, please describe how you have done so. What strategies did you use? What was the context (materials read, lesson topic, student activities)? What reading skills/knowledge did you focus on? Why those?
  ○ If not, talk about your thinking around the topic, i.e. why haven’t you? How important do you think it is for you to actively teach reading? Are there obstacles getting in the way, e.g. time, knowledge of appropriate strategies, other priorities?
• What gaps in your own knowledge about reading or instructional strategies specific to reading do you see? Are there any particular areas of assistance that you see you might need?
• Describe what you learned from the 2\textsuperscript{nd} round of reading assessments and the decisions you made regarding teaching of reading in your classroom.
• Identify two or three students and compare the results from the 1\textsuperscript{st} to the 2\textsuperscript{nd} administration of the reading assessments. Think about each student and how they may have interacted with your instructional activities. What do the results tell you about their reading skill development and the effectiveness of your teaching strategies for each student?
• Do you think that you have changed as a teacher in the way you approach reading since we started this process of focusing on reading assessment and instruction?
  ○ If so, identify two or three areas of change and describe what differences you have noted.
  ○ If not, why haven’t you? Do you feel that your level of expertise is roughly the same? If so, would you say that you already knew a lot about reading instruction, or that you haven’t had the time or the resources needed to expand your knowledge base or change your instructional practices?
• What are your plans for reading instruction moving forward now that your role in this study is concluded?
# Appendix 2: Record of Participation

<table>
<thead>
<tr>
<th>Teacher Name</th>
<th>Sch</th>
<th>Teacher Tests</th>
<th>Student Tests Admin 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ALI1&amp;2</td>
<td>MARS</td>
</tr>
<tr>
<td>Sandy</td>
<td>HS</td>
<td>X/X</td>
<td>X</td>
</tr>
<tr>
<td>Greg</td>
<td>HS</td>
<td>X/X</td>
<td>X</td>
</tr>
<tr>
<td>Albert</td>
<td>HS</td>
<td>X/X</td>
<td>X</td>
</tr>
<tr>
<td>Betsy</td>
<td>HS</td>
<td>X/X</td>
<td>X</td>
</tr>
<tr>
<td>Janis</td>
<td>MS</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gene</td>
<td>MS</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher Name</th>
<th>Student Tests Admin 2</th>
<th>PD</th>
<th>Interviews w/teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MARS</td>
<td>EDGE</td>
<td>TOS</td>
</tr>
<tr>
<td>Sandy</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Greg</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Albert</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Betsy</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Janis</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations**

- **ALI**: Assessment Literacy Inventory
- **TOS**: TOSCRF and TOSWRF (Test of Silent Contextual Reading Fluency and Test of Silent Word Reading Fluency)
- **MARS**: Metacognitive Awareness of Reading Skills Inventory
- **EDGE**: Lexile test from EDGE series – Hampton Brown/ National Geographic publisher
- **VOC**: Vocabulary Assessment – either VINE or textbook-based assessment
- **PD**: Professional Development on reading domain, assessments, instruction.
Appendix 3: Assessment Literacy Inventory (ALI)

Assessment Literacy Inventory

Cynthia Campbell, Ph.D. Northern Illinois University and
Craig A. Mertler, Ph.D. Bowling Green State University

Description of the ALI:
The Assessment Literacy Inventory (ALI) consists of five scenarios, each followed by seven questions. The items are related to the seven “Standards for Teacher Competence in the Educational Assessment of Students.” Some of the items are intended to measure general concepts related to testing and assessment, including the use of assessment activities for assigning student grades and communicating the results of assessments to students and parents; other items are related to knowledge of standardized testing, and the remaining items are related to classroom assessment.

Directions:
Read each scenario followed by each item carefully; select the response you think is the best one and mark your response on the answer sheet. Even if you are not sure of your choice, mark the response you believe to be the best.
Ms. O’Connor, a math teacher, questions how well her 10th grade students are able to apply what they have learned in class to situations encountered in their everyday lives. Although the teacher’s manual contains numerous items to test understanding of mathematical concepts, she is not convinced that giving a paper-and-pencil test is the best method for determining what she wants to know.

1. Based on the above scenario, the type of assessment that would best answer Ms. O’Connor’s question is called a/an
   A. performance assessment.
   B. authentic assessment.
   C. extended response assessment.
   D. standardized test.

2. In order to grade her students’ knowledge accurately and consistently, Ms. O’Connor would be well advised to
   A. identify criteria from the unit objectives and create a scoring rubric.
   B. develop a scoring rubric after getting a feel for what students can do.
   C. consider student performance on similar types of assignments.
   D. consult with experienced colleagues about criteria that has been used in the past.

3. To get a general impression of how well her students perform in mathematics in comparison to other 10th graders, Ms. O’Connor administers a standardized math test. This practice is acceptable only if
   A. the reliability of the standardized test does not exceed .60.
   B. the standardized test is administered individually to students.
   C. the content of the standardized test is well known to students.
   D. the comparison group is comprised of grade level peers.

4. Which of the following is an inappropriate use of the results from this standardized math test?
   A. planning instruction
   B. assigning student grades
   C. determining students’ strengths and weaknesses
   D. developing curriculum

5. Throughout instruction, Ms. O’Connor assesses how well her students are grasping the material. These assessments range from giving short quizzes following introduction to a new topic, to administering an end-of-the-unit final exam. In order to improve the validity of this grading procedure, Ms. O’Connor should
   A. make the grading scale the same for all assessments.
   B. consider students’ prior performance before assigning a final grade.
   C. weight assessments according to their relative importance.
   D. take into consideration each student’s effort when calculating grades.

6. During a parent teacher conference, one of the parents of a student in Ms. O’Connor’s class wants to know what it means that his daughter scored in the 80th percentile in mathematics. Which of the following provides the best explanation of this student’s score?
   A. She got 80% of the items on the math test correct.
   B. She is likely to earn a grade of ‘B’ in her math class.
   C. She is demonstrating above grade level performance in math.
   D. She scored the same or better than 80% of the norm group.

7. Which of the following is an appropriate use of assessment information?
   A. Utilize information from a variety of assessments when making decisions about student learning.
   B. Use scores from standardized tests to determine teacher instructional effectiveness.
   C. Use scores from a standardized test as the primary indicator of student retention.
   D. Post final grades in order to provide normative information to students in the class.
Appendix 4: Metacognitive Awareness of Reading Skills Inventory (MARSII)

Student name: ________________________ Teacher name: _________________________
Date: _________________________________

Directions: Listed below are statements about what people do when they read academic or school related materials such as textbooks or library books. Five numbers follow each statement (1, 2, 3, 4, 5), and each number means the following:

- 1 means “I never or almost never do this.”
- 2 means “I do this only occasionally.”
- 3 means “I sometimes do this” (50% of the time).
- 4 means “I usually do this.”
- 5 means “I always or almost always do this.”

After reading each statement, mark the number on the scanform (1, 2, 3, 4, or 5) that applies to you using the scale provided. Please note that there are no right or wrong answers to the statements in this inventory.

**Strategy**

<table>
<thead>
<tr>
<th>Statement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 I have a purpose in mind when I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 I take notes while reading to help me understand what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 I think about what I know to help me understand what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 I preview the text to see what it’s about before reading it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 When text becomes difficult, I read aloud to help me understand what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 I summarize what I read to reflect on important information in the text.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 I think about whether the content of the text fits my reading purpose.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 I read slowly but carefully to be sure I understand what I’m reading.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 I discuss what I read with others to check my understanding.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 I skim the text first by noting characteristics like length and organization.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 I try to get back on track when I lose concentration.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 I underline or circle information in the text to help me remember it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 I adjust my reading speed according to what I’m reading.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 I decide what to read closely and what to ignore.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 I use reference material such as a dictionary to help me understand what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 When the text becomes difficult, I pay closer attention to what I’m reading.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 I use tables, figures, and pictures in the text to increase my understanding.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 I stop from time to time and think about what I’m reading.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 I use context clues to help me better understand what I’m reading.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 I paraphrase (restate ideas in my own words) to better understand what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 I try to picture or visualize information to help me remember what I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 I use typographical aids like boldface and italics to identify key information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 I critically analyze and evaluate the information presented in the text.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 I go back and forth in the text to find relationships among ideas in it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 I check my understanding when I come across conflicting information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 I try to guess what the material is about when I read.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 When the text becomes difficult, I reread to increase my understanding.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 I ask myself questions I like to have answered in the text.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 I check to see whether my guesses about the text are right or wrong.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 I try to guess the meaning of unknown words or phrases.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix 5: Coding Configurations

<table>
<thead>
<tr>
<th>Code Name</th>
<th>Definition/Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedagogical knowledge: Instructional Strategies: Direct instruction</td>
<td>Modeling, read/think-alouds, reciprocal teaching</td>
</tr>
<tr>
<td>Pedagogical knowledge: Instructional Strategies: Indirect instruction</td>
<td>Independent projects/research, collaborative projects, jigsaw, etc.</td>
</tr>
<tr>
<td>Pedagogical knowledge: Engagement</td>
<td>Problem-based, HOTS, authenticity, inquiry, multiple intelligences</td>
</tr>
<tr>
<td>Contextual understanding: Purpose for reading</td>
<td>Testing, general interest, recreation, non-fiction, etc.</td>
</tr>
<tr>
<td>Contextual understanding: Student background</td>
<td>Parental support, SES, special population (SpEd, ELD), ethnicity, skill base</td>
</tr>
<tr>
<td>Contextual understanding: Learning environment</td>
<td>Classroom, relationships with students</td>
</tr>
<tr>
<td>Reading domain: Subject-area knowledge</td>
<td>Domain knowledge – breadth of one’s knowledge about a subject</td>
</tr>
<tr>
<td>Reading domain: Subject-area knowledge</td>
<td>Topic knowledge – about specific topic within a domain</td>
</tr>
<tr>
<td>Reading domain: Strategies</td>
<td>Cognitive (predicting, questioning, summarizing, etc.), metacognitive (self-awareness/monitoring)</td>
</tr>
<tr>
<td>Reading domain: Interest &amp; motivation</td>
<td>Situational/short-term – tests/projects, individual/long-term – disciplinary, involvement in field</td>
</tr>
<tr>
<td>Reading domain: content knowledge</td>
<td>5 pillars of reading: comprehension, phonemic awareness, vocabulary, phonics, fluency; narrative/expository,</td>
</tr>
<tr>
<td>Outcomes: Item Design</td>
<td>Performance-based, authentic</td>
</tr>
<tr>
<td>Outcomes: Data Collection</td>
<td>Observations, portfolio, data</td>
</tr>
<tr>
<td>Outcomes: Evaluation</td>
<td>Rubrics, scoring, grading</td>
</tr>
<tr>
<td>Assessment Literacy: Validity</td>
<td>Evidence, inference, interpretation</td>
</tr>
<tr>
<td>Assessment Literacy: Reliability</td>
<td>Benchmarks, common assessments, holistic scoring</td>
</tr>
<tr>
<td>Assessment Literacy: Construct</td>
<td>Framework, Progress/curriculum maps, criterion</td>
</tr>
<tr>
<td>Assessment Literacy: Strategies</td>
<td>Summative/formative, authenticity, performance-based</td>
</tr>
<tr>
<td>Assessment Literacy: Attitudes about assessments</td>
<td>Teacher motivation, creativity, value</td>
</tr>
<tr>
<td>Professional Development: Reading domain</td>
<td>Training completed, needed, conducted</td>
</tr>
<tr>
<td>Professional Development: Assessment Literacy</td>
<td>Training completed, needed, conducted</td>
</tr>
</tbody>
</table>
Reading Domain

Elements of Development (Alexander)
The reading domain is a developmental interplay between reader, text, and context with multiple stages of development.

*Subject-area Knowledge*
- Domain knowledge is breadth of one’s knowledge about a subject
- Topic knowledge is about specific topic within a domain

Interest
- Energizing learner’s underlying needs or desires
- Individual interest is long-term investment/involvement in field
- Situational interest related to temporary arousal triggered by conditions in immediate environment (i.e. test coming up)

Strategies
- Cognitive procedures – predicting, questioning, summarizing
- Metacognitive strategies – self-evaluation – related to monitoring one’s learning

Classes of Reading skill (O’Reilly)
1. Prerequisite – Decoding, recognition of printed text.
2. Model building – Construct meaning from decoded text, locate & retrieve information, & chunk, organize, summarize information. Helps with fast and efficient word identification skill foundation for text comprehension
3. Applied comprehension – Use information for particular purpose

Cognitive Skills Framework (O’Reilly)
1. Retrieve details – identify, locate, and retrieve details
2. Define Purpose – helps define goals and motivation
3. Organize information – ability to chunk, summarize and organize information
4. Ask deep questions – measure ability to ask and answer deep questions
5. Synthesize information – integrate and synthesize information from multiple related texts
6. Make inferences – Inference defined as any piece of info not explicitly stated in text. Measure quality of inferences and generalizations
7. Take a critical stance – adopt a stance critiquing the text with rationale and evidence

Habits of Good Readers (Valencia)
1. Demonstrate independence
2. Build strong content knowledge
3. Respond to varying demands of audience, task, purpose, and discipline
4. Comprehend as well as critique
5. Evaluate evidence
6. Use technology and digital media
7. Understand other perspectives and cultures

6 Reader Profiles
1. Highly competent – good linguistic knowledge, relevant schema, repertoire of strategies, interest in topic, “orchestrate” cognition and motivation to enrichment or satisfaction
2. Seriously challenged – Barriers include language processing, poor schema, few strategic skills, little motivation
3. Effortful processor – Lots of strategic effort in face of linguistic challenges or poor topic knowledge
4. Knowledge-reliant – Weak on strategic processing, dependent on existing knowledge
5. Non-strategic processor – Few strategies w/little self-monitoring
6. Resistant reader – Can have good knowledge and strategies but little motivation to put it all together
Appendix 7: Professional Development - Essential Research Findings about Effective Strategies for Teaching Reading

<table>
<thead>
<tr>
<th>Cognitive Strategies</th>
<th>Network Connections</th>
</tr>
</thead>
</table>
| • *Balanced classroom instruction* includes both explicit instruction in specific comprehension strategies and time for actual reading, writing, and discussion of text.  
  • *Think-alouds* valuable for helping students become more metacognitive in improving own reading abilities. Recognizing when struggling and what helps get them through problems. | • Perfetti's lexical quality hypothesis (Perfetti & Hart, 2002) asserts quality of *abundance of networks* of associations w/a word is a correlate of comprehension of text.  
  • *Connecting* fictional experiences to *students’ lives* through inquiry projects, front-loading of information |

<table>
<thead>
<tr>
<th>Vocabulary</th>
</tr>
</thead>
</table>
| • Word learning should be *multidimensional* including not only linguistic dimensions of a word, such as etymological breakdown, but also visual, auditory, or other sensory images.  
  • *Morphological awareness* found to have significant positive impact on comprehension of complex words in 8th & 9th grades.  
  • Students acquire vocab best when used in *meaningful, authentic contexts*, accdg to NRP review of 50 studies published in refereed journals  
  • Words are pretested before the unit, posted on content area word wall, *deliberately taught*, used by student and teacher multiple times and post-tested at conclusion of unit. |

<table>
<thead>
<tr>
<th>Effective Strategies</th>
</tr>
</thead>
</table>
| • *Associating* - form word-net framework to make connections and determine word relationships  
  • *Contextualizing* - allows readers to increase vocabulary & comprehension by using context that surrounds unknown word  
  • *Visualizing* - students create image representing meaning of word or concept  
  • *Personalizing* - increases sense of ownership of a word. Need definition and context as well as multiple exposures  
  • *Referencing* - allows readers to use resources to determine meaning of a word when unknown  
  • *Literature circles* in which choice primary component, particularly useful if tied to theme being studied, as WW 2. |
Bibliography


