

A Study of High Quality Teaching: Mathematics and Reading

by

Linda Valli
Robert Croninger
Patricia Alexander
Marilyn Chambliss
Anna Graeber
Jeremy Price
University of Maryland

Paper Presented at the Presidential Invited Symposium, *Looking in Classrooms: Again*, Annual Meeting of the American Educational Research Association San Diego, CA, April 15, 2004. Direct all correspondence to Linda Valli at LRV@ umd.edu. The work reported herein was supported by the Interdisciplinary Educational Research Initiative (IERI # 0115389), a combined effort of the National Science Foundation, the U.S. Department of Education, and the National Institutes of Health. The opinions expressed in this manuscript are our own and do not reflect the positions and policy of the National Science Foundation, the U.S. Department of Education, or the National Institutes of Health.

A Study of High Quality Teaching: Mathematics and Reading

The High Quality Teaching (HQT) study is based on the now widely-accepted notion that teachers have a significant influence on student learning (NCTAF, 1996; Sanders & Horn, 1998). Recent studies indicate that teacher preparation and experience, as well as investments in teacher learning, substantively contribute to student learning gains and account for considerable variance in student achievement (Cohen & Hill, 2000; Ferguson, 1998; Ferguson & Ladd, 1996; Greenwald, Hedges, & Laine, 1996). These studies provide compelling evidence that qualified teachers make more of a difference than past research (Coleman, 1966; Jencks, 1972) suggested was possible and that teachers might make more of a difference than structural or programmatic educational reforms (Hawley & Rosenholtz, 1984). This body of research is, however, quite limited in measures of actual classroom practice. By using only global indicators of *teacher* quality (years of experience, level of education, licensure status), often called presage variables (Shulman, 1986), many of the studies are silent on the issue of *teaching* quality.

A large body of research, beginning with the early process-product studies (Dunkin & Biddle, 1974), has examined the relationship between teaching process variables and student outcomes. This literature offers some broad constructs of teaching practices as well as methodological strategies for how to investigate teaching, but has been criticized for being theoretically underdeveloped, needing more refined measures, ignoring context variables, and being too general to offer meaningful guidance to teachers (Brophy & Good, 1986; Floden, 2001; Shulman, 1986). Because of their small and unstable effects, these generic measures of teacher behaviors have also been criticized on the basis of construct validity (Rowan, 2000). We still need to know much more specifically what good teachers do on a regular basis, during the normal course of teaching, to promote student learning, apart from external reform efforts,

and how this teaching compares across subject areas.

But how to study teaching, and the quality of teaching, are themselves matters of contention. Historically, two fundamentally different traditions have informed research on teaching: the descriptive and the prescriptive. Within the descriptive tradition, researchers look inside classrooms to understand better what teachers and students do. The quest is for basic understanding of classroom life apart from judgments of what ought to happen there (Jackson, 1968; Kliebard, 1973; Lampert, 2001). Descriptive researchers claim that prescriptions for teaching often fail because they do not adequately consider the complexities of classroom teaching and argue for more finely-tuned descriptions of different classroom contexts.

In the prescriptive tradition, the impulse is not merely to understand and describe classroom life, but to improve it. This requires knowing what to prescribe. But where do prescriptions for good teaching come from? How does one know what constitutes high quality teaching? What are the sources of that knowledge? Prescriptive researchers part ways on these questions, sometimes quite dramatically. One group claims that images of good teaching should be derived from expert opinion—from standards, norms, theories, and nominations of good teachers. These experts might have different disciplinary groundings, views of learning, or social goals that shape their conceptualizations, but they all begin from a priori judgments. Shulman (1996) uses the word “correspondence” to describe this research approach in which “a given exemplar of instruction is compared to a model or conception of good teaching derived from a theory or ideology” (p. 28). A problem with this tradition is that, as the reading and math wars so vividly portray, agreement on any given set of standards is hard to achieve.

In contrast to the correspondence approach, another group of researchers looks, not to standards *for* teaching, but to the consequences *of* teaching. Teachers’ work is considered to be

good if it produces desired outcomes, usually students' academic learning. Shulman (1996) calls this a "pragmatic" or "correlative" conception of teaching effectiveness. In this conception, teaching and learning are closely connected, with students' standardized test scores used as the measure of high quality teaching. But, as its critics charge, the tests themselves could assess a very narrow range of learning, and achievement gains—as measured by these tests—could result from practices antithetical to most conceptions of good teaching, practices such as teaching to the test, narrowing the curriculum, or manipulating the population of test takers (Darling-Hammond, 1997; McNeil, 2000).

These distinctions in research traditions are important because they point to underlying dilemmas in designing studies of teaching and deciding what constitutes evidence of high quality teaching. Is there still value in gathering purely descriptive data on life in classrooms? Should a teaching practice be considered high quality because it conforms to a priori judgments or because it is associated with measures of student learning? Must it do both? Researchers have begun to use a variety of strategies to link descriptive, normative and correlative indicators of teaching. But going from these separate traditions to theoretically-based, relational approaches raises a complex set of research challenges. We discuss some of these challenges by describing our current research in elementary classrooms.

Overview of the HQT Study

High Quality Teaching is a four-year study that focuses on what teachers do to help 4th and 5th grade students succeed in reading and mathematics, as well as on how various educational policies and organizational factors influence the ability of teachers to scale up and sustain good teaching practices over time.

Research Questions

Our goal is to learn more about what teachers and programs do to assist students who are struggling to acquire foundational skills by answering five research questions:

- What do highly successful 4th and 5th grade teachers do to help students achieve above predicted rates in the areas of mathematics and reading?
- What do these teachers do to expand learning opportunities for all students and help close the achievement gap in foundational skills between traditionally high and low performing groups?
- Do teachers change their pedagogical practices over time? If so, what is the nature of those changes? Are documented changes in teachers' practices associated with shifts in their learning or that of their students?
- What, if any, educational policies and instructional programs influence teaching practice?
- What is the correspondence between our constructs of high quality teaching and student achievement?

We selected 4th and 5th grade reading and mathematics because these are arguably the two most important subjects in the elementary school curriculum, yet many 4th and 5th graders still struggle with foundational literacy and numeracy skills. Thirty-seven percent of 4th graders nationally scored below the basic level of reading on the 2003 NAEP and 68% scored below the proficient level (<http://nces.ed.gov/nationsreportcard/>, March 27, 2004). In the 8th grade, 26% are still below the basic level. Some groups of children are particularly vulnerable with regards to literacy, with Black, Hispanic, and American Indian children scoring lower, on average, than White students (Donahue, Voelkl, Campbell, & Mazzeo, 1999), a pattern that is also found in international comparisons (National Research Council [NRC], 1998). In mathematics, results are similar. Despite the increase in NAEP scores at every grade level tested from 1990 to 2003, the latest NAEP Mathematics Assessment results indicate that almost one out of four 4th graders and one out of three 8th graders are still below the basic level (<http://nces.ed.gov/nationsreportcard/>, March 27, 2004). These results suggest that a significant portion of upper elementary school

students have not achieved mastery of fundamental skills and concepts.

Research on 4th and 5th grade mathematics and reading is greatly needed. The teacher effectiveness research of the '70s and '80s is too generic to offer much guidance, and the subject specific research generally focuses on other grade levels. In the area of mathematics, recent studies tend to look at early primary grade levels (e.g., Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Cobb, Wood, Yackel, Nicholls, Wheatley, Trigatti, & Perlwitz, 1991) or secondary school grade levels (e.g., Schoen, Hirsch & Ziebarth, 1998). Most of these studies have as a primary concern the degree of implementation of a specific mathematics program as it relates to student achievement, not teaching practice per se. In the area of reading, recent studies in both primary (Taylor, Pearson, Clark, & Walpole, 1999; Pressley, Rankin, & Yokoi 1996) and fifth grades (Pressley, Yokoi, Wharton-McDonald, & Mistretta, 1997) have identified effective instructional practices that are relevant for struggling learners. However, the one fifth grade study surveyed rather than observed and interviewed teachers, and the frameworks for all three studies were more empirically than theoretically based. Furthermore, although many of the same adults teach both reading and mathematics, there are surprisingly few comparative studies of the teaching of reading and mathematics by the same teachers (Chambliss & Graeber, 2003).

Site Selection

The sites for the High Quality Teaching study are elementary schools in the Montgomery County Public Schools (MCPS) system. Located in the state of Maryland, MCPS is one of the largest and most diverse school districts in the nation. The district enrolls more than 130,000 students and is less than 50% White, after having been more than 90% White in the early 1970s. More than 30% of students are on Free and Reduced-Price Meal Services (FARMS) and 20% are currently, or have been, enrolled in English for Speakers of Other Languages (ESOL)

programs. With 7,400 students who speak 119 languages, Montgomery County represents more than half of all students in ESOL programs in the entire state.

Because of our interest in teacher success with low-achieving students, we selected schools with moderate to high levels of poverty and higher than expected levels of student achievement. MCPS provided us with four years of average mathematics and reading scores from Criterion Reference Tests (CRTs) for 3rd, 4th, and 5th grade students at 119 of its elementary schools. These data enabled us to investigate average achievement given free/reduced price enrollments and annual changes in achievement between 1997 and 1999. We used two different strategies to identify schools: (a) OLS regression to estimate the average effect of Free and Reduced-priced Meals (FARMS) enrollments on average CRT mathematics scores and average CRT reading scores, and (b) hierarchical linear modeling to look at changes in achievement between 1997 and 1999 after controlling for free/reduced-price lunch enrollments. HLM indicated that growth or change varied significantly between schools, with some schools realizing greater improvements in scores in both mathematics and reading than other schools.

To expand the number of schools in the second year of the study, we analyzed MCPS data on 4th grade CTBS achievement scores (2002) and invited additional schools to participate that met three criteria: They were higher than the district average on percentage of FARMS students, had better than predicted average achievement scores controlling for FARMS enrollment, and had a lower than average achievement gap between FARMS and non-FARMS students. In the third year of the study, we supplemented these analytic strategies with nominations from knowledgeable colleagues. Once schools were selected, they received a formal invitation to participate from MCPS. This letter was followed by an MCPS phone call and a visit to the principal by HQT team members. HQT staff then met with teachers to

explain the study and obtain their consent. During the first year, 67 teachers in 11 schools participated; during the second year, 73 in 16 schools; and in the current, third year, of the study, our pool has expanded to 18 schools and 77 teachers.¹

How is the HQT Study Conceptualized?

Our conceptualization of high quality teaching is based on a broad review of literature across various fields of education research: literacy, mathematics education, cognitive psychology, and sociology. This review tells us that HQT is a complex, multi-dimensional phenomenon that is best studied through a variety of over-lapping, complementary strategies. The general contours for current theories and models of teaching are increasingly set by core principles about how children learn, and, correspondingly, how successful teachers teach (Croninger, Valli, & Price, 2003). What distinguishes these beliefs from earlier propositions about teaching is that they are based on an emergent body of research about learning (Alexander & Murphy, 1998; Bransford, Brown, & Cocking 1999) that is linked to rich descriptions of successful practice (Lampert, 2001; Ball & Cohen, 1999). In our own study of high-quality teaching, we draw on five research domains that were core to the development of the 14 Learner-Centered Psychological Principles (American Psychological Association [APA] Board of Educational Affairs, 1995).

But differences in subject areas, topics, students, and curricular materials make different demands on teachers' content and pedagogical expertise, signaling the importance of drawing on domain-specific research (Alexander & Fives, 2000; Croninger, Valli, & Price, 2003; Shulman,

¹ Because of turn-over of schools and teachers across the three years the number of teachers did not increase dramatically, even though we added new schools and teachers each year. The rapid turn-over of teachers also meant that the student achievement data used to select schools were outdated. Many of the teachers responsible for the above-predicted student learning were no longer at the participating schools. That is why we shifted to collegial nominations to select Year 3 schools.

1986). While this research is consistent with the learner-centered psychological principles, the particular manifestations of exemplary teaching differ in reading and mathematics. They would also differ according to the different goals and values that drive the instruction (Chambliss & Graeber, 2003). As is evident from the "Reading and Mathematics Wars," the meaning of high-quality teaching is still hotly debated.

In general, however, much of the research on exemplary reading instruction at the 4th and 5th grade levels suggests the importance of large amounts of reading and writing, connections across the curriculum, and skills instruction imbedded in meaningful interaction with text. According to this body of research (Morrow, Wamsley, Duhammel, & Fittipaldi, 2002; Pressley, Yokoi, Wharton-McDonald, & Mistretta, 1997; Taylor, Pearson, Clark, & Walpole, 2002), students would have choice about interesting and challenging texts of a variety of genres, and would be encouraged to respond personally to those texts. Teachers would promote dialogue about reading and writing through teacher questioning, grouping practices, cooperative learning, and informative feedback. In mathematics, high quality teaching at the 4th and 5th grade levels would engage students in high-level tasks and maintain the level of tasks while providing scaffolding as needed. Teachers would question students' thinking, be responsive to that thinking, and would make alternative methods of solving problems a focus of classroom discussions. Students would be encouraged and helped to make connections between and among mathematical ideas and to see the usefulness of mathematics outside the classroom (Fennema et al., 1996; Hiebert & Lefevre, 1986; Kilpatric, Swafford & Findell, 2001; NCTM, 2000; Shafer & Romberg, 1999).

But understanding high-quality teaching requires not only understanding what teachers know and do, but understanding the organizational and normative contexts in

which they work. Classrooms can be structured in ways that promote or undermine both opportunities and motivation to learn. Outside the classroom, the school's organizational capacity and external support (particularly school district support) affect instructional quality and learning opportunities (Newmann & Wehlage, 1995; Valli, 2000). The HQT study looks beyond classrooms and teachers as the unit of analysis to examine ways in which schools organize themselves, attract resources to support teaching and learning, and negotiate external mandates. It theorizes classrooms as nested within loosely coupled organizations, or what Sarason (1999) has called "markedly uncoordinated" systems that have enormous impact on teachers' performance and sense of professional well being. While there is no uniform set of principles about how organizational designs influence teaching, key factors are the ways in which district and school-level administrators mediate messages about teaching and learning, value teachers' judgments and knowledge, distribute resources, provide opportunities for teachers' on-going learning, and assess student learning (Croninger, Valli, & Price, 2003).

How is the HQT Study Operationalized?

In keeping with this multi-faceted conceptualization of teaching, we use multiple data gathering instruments to understand that phenomenon better and to increase the validity of our findings. We use *quantitative* methods to identify case sites, summarize a wide range of data about teachers and schools, and test multilevel models of how schools and teachers affect student achievement in reading and mathematics. We use *qualitative* methods to refine data collection instruments and models of teaching quality, render thick descriptions of high quality teaching, explore participant perspectives, and gather school and district-level data. These two complementary lines of inquiry allow us to create an unusually detailed, longitudinal data set appropriate for investigating how teaching quality influences student learning in reading

and mathematics.

We describe below our four main types of data collection strategies and instruments (time sampling, attribution scales, daily logs, and case studies/interviews) and the trade-offs that occur by our methodological choices. Does our mixed methods model provide a rich, comprehensive picture of high quality teaching or a disjointed, fragmented picture? What are the challenges in trying to relate data from instruments that conceptualize data in very different ways, based on different assumptions and epistemologies? Although each of these strategies has its own theoretical base, we see each as only a partial window on good teaching. But how does one look across theoretical frameworks? Are our ways of looking always complementary and reinforcing or do they create interpretive nightmares?

Time Sampling Observation Protocols

In order to capture teaching profiles, researchers have long relied on time sampling, which represents instructional practices as a series of sequentially-related actions and decisions (Good & Brophy, 2003). The High-Quality Teaching Study is no different. With an anticipated sample size of approximately 120 classrooms, one of our first decisions was to use standardized instruments for frequency counts of regularly occurring classroom behaviors. A second decision was to use *time* sampling rather than *event* sampling. Unlike event sampling, which pre-determines and records only behaviors of interest when they occur, time sampling codes snapshot samples of teachers' normal behaviors at regular intervals, providing a more inclusive and representative record of regularly occurring behaviors throughout the class period. This choice is consistent with our theoretical perspective. Although we came to the study with theory-based notions of high-quality teaching, we wanted to maintain openness to various manifestations and not presume or impose one particular model. In keeping with the

descriptive tradition of research on teaching, we sought to learn more about the daily texture of reading and mathematics lessons.

The project developed two observation protocols, one for reading and one for mathematics and programmed the protocols on laptops using AccessTM. The computer screen prompts the HQT observer every three minutes to enter data in seven major categories: Teacher Activity, Student Activity, Classroom Organization, Content, Context, Classroom Behavior, and Technology/Materials. Developed through extensive reviews of the literature, pilot tested, and revised to reflect both research and classroom realities, these protocols enable us to see if the frequency data produce similar or dissimilar profiles of teaching practices. Since we will have six-eight lessons of approximately 20 coding episodes per lesson over multiple years, we are able to create highly-detailed profiles of reading and mathematics classes.

But we fully realize that the classroom “realities” portrayed by the time sampling instrument (or any instruments) are the product of the methodological decisions researchers make, an artifact of constructed categories, items, and decision rules. These categories, items, and rules provide certain information about the classroom, obscure other information, and create specific types of coding dilemmas. Take just one example, decisions about the grain-size of items in reading and mathematics, that has affected the way in which we gather data on the subject matter of the lesson, what we call “Episode Content.”

Since the comparison of mathematics and reading instruction was of primary interest, our decision rule was to use generic categories whenever possible, but to depart from that rule when important information about subject matter would be lost. This was most obviously the case with Episode Content. In reading, in keeping with the domain-specific research, we chose to record at a smaller grain size than in mathematics. Within the broad category of reading, we differentiate,

for example, among comprehension, strategy, and fluency instruction. Within comprehension, we further focus on genre, theme/main idea, story elements/text design, personal response and literal response. Within the broad category of writing, we differentiate between writing related to reading and writing unrelated to reading, with specific items under each of those categories. In addition, we have a series of items that code instruction that is independent of the reading and writing of coherent text (e.g., vocabulary, decoding, spelling, conventions).

In contrast to these relatively specific, concrete literacy categories, we chose to record the mathematics content within the broad, more abstract categories of Conceptual, Procedural, and Linking Conceptual and Procedural. This “division” of lesson content is quite different from the more topical divisions (reading, writing, vocabulary) in the reading instrument. Two factors drove this decision. First, teachers in the study enter data on the content covered in their daily logs and this level of reporting typically has a high degree of accuracy (Porter, 2004; Rowan, 2004). Second, numerous studies already examine specific ways to teach a specific outcome. We were interested in examining a knotty issue at the heart of the on-going math wars debate, that is, the relative significance of conceptual versus procedural, understanding versus skill, or relational versus instrumental understanding in mathematics education (Hiebert, 1986; Skemp, 1987).

But gathering data at different grain-sizes makes comparative analysis difficult and creates different kinds of data collection problems. In reading, episode content created some interrater reliability problems due to the sheer volume of items from which to choose. Reading instruction can shift from theme to story elements to literal response to vocabulary within seconds; it can also shift across reading and writing quickly as teachers work to integrate those literacy areas. In mathematics, coding problems were sometimes deeper, more a question of

construct validity and stability. With only three main content categories to code, one would expect interrater reliability to be higher and it is (*confirm*). But although the differences among procedural, conceptual, and linking procedural/conceptual mathematics might seem clear and obvious, the categories proved to be more ambiguous and overlapping than we had anticipated. The intent of a teacher question or the nature of the students' thinking that was prompted by the question wasn't always clear to us as observers.²

To attend to these and other types of validity and reliability issues, the observation protocols needed on-going refinement and clarification. As much as possible, we tried to make items concrete and low inference, identifying behaviors that were readily observable while preserving meaningful chunks of classroom interactions. Observers were required to participate in intensive orientation, training, and testing before they collected data on their own. For coding consistency, the faculty and staff produced detailed observer glossaries and classroom scenarios,

² Consider one simple example: a 4th grade review lesson on factors and multiplication. After briefly revisiting the meaning of "factor" and eliciting some examples, the teacher gives her students a work sheet that asked them to list all the factors for 6, 12, 18, 24, and 20; to select the multiple choice item that listed all the factors of 28 and of 16; and to show their work. Students busily get to their work. But what mental task engages them? Is their understanding of "factor" what enables them to quickly write the numbers 1, 2, 3, and 6 as factors of 6? If so, the episode content is conceptual. Or are they translating the questions into rote multiplication problems—a procedural task—to see which combinations produce the specified numbers? Or does the task require them to make explicit connections between a procedure ($2 \times 3 = 6$) and a concept (factor)? Are they linking their conceptual and procedural knowledge as they answer the questions? And if explicit links between this concept and procedure are required for the first questions, does this remain the case when each question follows the same pattern? Or have they proceduralized their understanding? Has the repetition enabled them to stop thinking conceptually and simply perform a mathematical operation?

In addition to the problem of reading students' minds, observers can have difficulty with teachers' questions and actions that can be ambiguous—read in different ways, have multiple functions or interpretations. Some students might construe a question to be asking for steps needed to solve a problem; other students may be attempting to recall of a list of needed symbols or forms. Furthermore, any procedural action necessarily has a conceptual component. One cannot engage in a content-free procedure. Therefore, the observer is always making a judgment about the primary emphasis or intent of teachers or cognitive work of students. Are their minds primarily focused on the steps in carrying out a problem (even a story problem), the meaning of a mathematical concept or formula, or the relationship between the steps and the ideas?

and continue to hold bi-weekly training sessions to discuss coding questions. “Experts” pair with observers throughout the year to prevent observer drift and idiosyncratic coding patterns. Staff compute interrater reliability scores to identify and target areas in need of rule reminders or concept clarification.

While descriptive statistics of individual observation items are of interest in themselves, they are just at the beginning of data analysis. In addition to item analysis, we are engaged in construct analysis: identifying key constructs from the literature and examining the extent to which they exist in our data. We have, for example, worked on a construct of high-quality interactions from items in the Teacher Activity/Student Activity categories, such teacher requesting student self assessment and elaboration, and students responding with conjectures, explanations, or alternative methods or answers. We are also looking at items that tap into the constructs of cognitive-demand, direct instruction, reform instruction, and classroom discourse. But there are analytic challenges in transforming discrete, nominal behavior items into theoretically-based constructs and in finding ways to assign numbers to these measures.

Attribution Scales

At the conclusion of each time sampling period, observers complete the Attribution Scale. The Attribution Scale was developed to capture the gestalt of a class session in terms of the relative demonstration of actions deemed important to effective learning within the psychological literature (Alexander & Murphy, 1998). As such, this more high-inference measure was seen as a complement to the Time Sampling protocol. Specifically, the items for the Attribution Scale were drawn from five domains of research that Alexander and Murphy (1998)

found to be core to Learner-Centered Psychological Principles (APA Board of Educational Affairs, 1995). Those five domains and the guiding premises central to effective learning are:

- *Knowledge base.* One's existing knowledge serves as the foundation of all future learning by guiding organization and representations, by serving as a basis of association with new information, and by coloring and filtering all new experiences.
- *Strategic processing or executive control.* The ability to reflect on and regulate one's own thoughts and behaviors is essential to learning and development.
- *Motivation and affect.* Motivational or affective factors, such as intrinsic motivation, attributions for learning or personal goals, along with the motivational characteristics of learning tasks, play a significant role in the learning process.
- *Development and individual differences.* Learning, although ultimately a unique adventure for all, progresses through various common stages of development influenced by both inherited and experiential or environmental factors.
- *Situation or context.* Learning is as much a socially shared undertaking as it is an individually constructed enterprise (Alexander & Murphy, 1998, p. 26).

The Attribution Scale contains four items for each of these five domains. For example, the Knowledge domain includes the item "Promotes principled understanding in students;" while the Strategic Processing domain asks whether the teacher "Models general or domain-specific strategies." Observers rate the teaching on a Likert-type scale ranging from 1 (*Not Evident*) to 4 (*Pervasive*). In addition to these separate item ratings, the observer assigns an overall rating to the instruction ranging from 1 (*Low Quality*) to 4 (*High Quality*).

Like all high-inference instruments, the Attribution Scale poses a particular set of epistemological and methodological challenges. Each classroom observer comes to the task with tacit or explicit notions about what constitutes more or less effective teaching. Even with training that offers a range of models, the observers are likely to retain basic beliefs about effective or ineffective pedagogy that cannot be readily altered. Just as with their beliefs about the nature of quality teaching, raters are presumed to have personal theories of learning under which they

operate. Those beliefs, whether tacit or explicit, can color their responses to scale items.

Beyond these general beliefs about teaching and learning, raters may well possess certain beliefs about reading and mathematics that enter into their judgments of quality. They may have certain expectations as to what should or should not be observed during reading or mathematics class—domain-specific expectations that affect their judgments. For instance, those who see more holistic instruction as preferable to explicit skills instruction in reading are apt to carry those domain-specific beliefs into the observations and, thus, the attributional ratings. Further, judgments about the level at which particular components are exhibited can be significantly influenced by the raters instructional experiences in classroom comparable to those being observed. Those with minimal classroom experience have fewer models against which to make comparative judgments, while those who are deeply immersed in the educational culture may be less able to view novel lessons from an alternative perspective.

Over the course of the project, observers can also begin to form more or less positive opinions of individual teachers that can color their rating accuracy. Rating a teacher a 1 or 4 on particular items in two consecutive visits might pre-condition an observer to rate the next lesson the same, even though the nature, delivery, and teacher’s understanding of that lesson might be quite different from the previous lessons. Or watching a mathematics lesson where several attributes were not evident might pre-condition an observer to over-estimate their prevalence in the reading lesson that is observed immediately after. And although we assume that the five dimensions measured on the attribution scale underlie effective teaching, we do not assume that the specific components would appear consistently from lesson to lesson. That is why the scale is a frequency measure—“not evident” to “pervasive”—rather than a quality measure such as “poor” to “outstanding.” Observers must be trained to think of lessons independent of one

another and not to think that giving a “1” on a particular attribute necessarily indicates a weak teacher or a weak lesson.

The research team employed several strategies to help observers establish a common frame-of-reference for rating purposes and overcome some of the rating problems associated with high-inference instruments. As with the time-sampling protocol, they prepared a detailed glossary that includes extended definitions and supporting examples for both reading and mathematics. These glossaries are part of the orientation training manual and can be accessed electronically from the computer protocols during the coding. Observers complete training sessions where taped lessons are viewed, scored, and interrater agreement is calculated. An “expert” coder also joins observers on classroom visits. Interrater agreement is calculated from these joint visits and questions or coding discrepancies are discussed on the spot and, later, in group meetings.

The collection of attribution data across multiple lessons and multiple years in two subject areas allows us to look for strong and consistent patterns in teachers’ attention to the dimensions of knowledge, strategies, motivation, individual differences, and the learning context. These attributional data do not sit alone in the project design, but serve as one important piece of a complex puzzle. Thus, attributional ratings can be compared and contrasted to other low and high inference indicators of teaching quality. We are curious to see whether the attribution ratings will validate the time-sampling observation profiles and vice versa.

Teacher Daily Logs

While the time-sampling instruments gather data on instructional practices and the attribution scale renders an overall gestalt of lessons, the teacher logs measure curriculum coverage and its distribution across students within a classroom over an entire year.

Participating teachers keep daily logs detailing their curriculum and tracking the activities of one of their students each day. Originally intended as simple paper and pencil checklists of coverage, depth of coverage, pace, and alignment of coverage and student assessments, these measures were to provide substantial information about the use of class time.

We soon realized, however, that having teachers collect logs on Personal Digital Assistants (PDAs), instead of paper/pencil, would have major advantages. PDAs would allow us to gather more complex data, facilitate the creation and maintenance of a comprehensive database, and be an incentive to participate (i.e., teachers could keep the PDAs and use them for other purposes). The power of software allowed us to branch into specific subcategories and collect continuous as well as categorical data. In addition to the original four measures we intended to collect, we now have measures of classroom organization, ability grouping and differentiation, student grade-level performance, and technology use. We also have more refined measures of curriculum coverage than originally thought possible.

To identify the content for the log protocols, HQT staff reviewed research, collected the curriculum guidelines for Montgomery County, and conducted a pilot test with teachers in a participating school. Consultations with district-level curriculum specialists helped us ensure content validity. Even though our study centers exclusively on 4th and 5th grades, we included curriculum topics that spanned 3rd – 6th grades and reflected the school district's new curriculum frameworks that were still in the development stage. Pilot teachers told us that differentiation of instruction within their classrooms would make class-level answers invalid. Therefore, we decided to collect data at the individual student level rather than the class level, one of our most important decisions. These individual-level data enable us to determine teacher differentiation of practice by their beliefs about students' prior performance and knowledge. Once teachers

give us their class lists, we use a web-based rostering program to download names and maintain data files that link students and teachers by subject.

Still designed as a checklist, the protocols walk teachers through a sequence of inter-related fields that are closely aligned with the Classroom Organization, Content, and Technology/Materials categories of the time-sampling observation instrument. This alignment allows us to check coding reliability by comparing log entries with observations conducted on those days. Although the checklist format makes the protocols easy to use, some of our teachers had never held a PDA before. So in addition to training sessions on the protocols themselves, we conducted training on PDA operation, maintenance, and other programs, such as the calendar, address book, to do list, and memo pad. Our goal was to help teachers become comfortable with PDAs and want to use them on a regular basis. We have prepared paper glossaries for teachers that describe each field in their reading and mathematics protocols and have a rapid response system in place for content questions and technical problems. By installing remote site software in each school, teachers download log records to the HQT server. This enables us to produce summary reports of log entries so individual teachers can compare their classroom profiles to that of the entire sample. Some of our analytic challenges are partitioning the student data base to examine differentiated instruction, matching items to relevant constructs, and comparing the sub-set of log data with time-sampling data.

Case Studies and Interviews

Our qualitative data, in the form of interviews and detailed field notes, provide a rich context in which these quantitative data can be better understood. Teachers' and principals' perspectives, their intentions and meanings, are an invaluable lens for interpreting patterns of teaching. The qualitative data include teacher interviews about observed lessons, in-depth

case studies of classroom teaching, focus group interviews, and principal interviews.

Time-Sampling Interviews and Artifacts. For each classroom observation, we try to conduct brief interviews where we obtain the teachers' constructions of the lesson: what was their intended lesson and what were their reflections on the enacted lesson. By field testing the interview protocols, we were able to eliminate redundant and extraneous questions and thus reduce the time burden on teachers. The core questions we ask before the lesson are: What is the main goal of the lesson; how does it fit into an overall unit? Where did you get your ideas? Is this something you've done before? What resources support the lesson? Post-observation questions are: Did the lesson go as intended? Where will you go from here? Teachers have these questions in advance, and have the option of writing answers before the observation take place. As part of the computer software, the interview protocols provide a database of teacher thinking and planning. During the time sampling coding, observers can keep notes on unusual occurrences that provide a coding context and become a permanent part of the database. We also ask teachers for copies of materials students are given. These artifacts help us flesh out our understandings and recollections, and give us a more concrete picture of curriculum coverage and teaching practices.

Case Studies. In addition to the qualitative data collected during time-sampling observations, we simultaneously audio-tape and take running field notes of selected classes. These case studies render thick descriptions of mathematics and reading lessons that represent an array of high quality teaching practices. In order to select cases, the first question we had to answer was, "What do we want cases of?" Because we are studying *teaching*, and not *teachers*, we cast the net broadly to capture a wide variety of lessons. Case study teachers are selected because they exhibit practices judged to be exemplary by observers, score high on the

attribution scale, and/or produce student learning at rates higher than predicted by students' prior achievement and FARMS status. Data collection for the case studies attends to the classroom environment/community; teacher/student discourse, subject-matter representation, learning tasks/materials/assessments, and student work samples. Case study teachers are interviewed after each observation to give us deeper understanding of their teaching beliefs, knowledge, and expectations.

Focus Group Interviews. To supplement these in-depth cases we held focus group interviews of teachers across participating schools. So far we have conducted four focus group meetings, interviewing 14 teachers from 12 of the 16 project schools in order to gather participant perspectives on high quality teaching, the supports and impediments to such teaching, and the impact of recent school district changes on their teaching. In addition to representing most of our participating schools, focus groups represented both subject areas, both grade levels, and the full range of age, gender and experience within our population of participating teachers. During data analysis, we keep four questions in mind: (1) How do focus group teachers describe high quality teaching, (2) what are its characteristics, (3) how is it supported, and (4) how is it constrained? Because the focus group teachers discuss both mathematics and reading, we are careful to distinguish between statements made about instruction in the two content areas. At this time, some of our main coding categories are teaching philosophies, values, goals, strategies, resources, and outcomes. All comments are treated as representations of personal theories and belief systems about high quality teaching

Principal Interviews. Standardized protocols were developed to interview principals about the school schedule, teacher/student assignments, resources, the curriculum, special programs, and funding. The purpose of these interviews is to collect information about the

school context and the principal's perceived role in relation to 4th and 5th grade mathematics and reading instruction. An interview is held with each of the principals in participating schools every year. In follow-up interviews, we directly solicit responses to our central research questions: "What are their perspectives on high quality teaching practices? How do they attempt to close the achievement gap? What school policies do they view as most effective in these efforts? Do they see teachers' practices changing over time? How and Why?" Like most of the qualitative data collected in our study, these interviews are audio taped, transcribed, stored, and coded in NVIVOTM. Hierarchical trees of coding categories were developed from the coding group's collaborative analysis of one principal interview. Some of the main coding categories (parent trees) for principal interviews are: school change, leadership, staffing, scheduling, state testing, resources, curriculum, and goal setting. Team members worked in pairs during subsequent coding of principal interviews to add to coding trees and refine their organization.

Our mixed methods research design of quantitative and qualitative methods presents some challenges, particularly in comparing and coding across deductive and inductive approaches. In analyzing the qualitative data, we consulted the time-sampling and attribution protocols as we encountered examples of teacher activity, student activity, lesson content, and so forth. Using the quantitative tools to help us formulate coding categories, we have a common starting place for analyzing the data from the different teachers. However, although we eventually want to link the qualitative data with the quantitative data, our primary goal is to preserve the meaning that individual teachers ascribe to what they did in their classes and the ways they speak about their personal teaching beliefs. Therefore, coding categories use the language of the teachers and are more descriptive than the preconceived constructs of the quantitative tools.

What Have We Learned About the Study of High Quality Teaching?

In this final section of the paper, we discuss what we have learned so far from our study of high-quality teaching. Because we are in the initial stages of data analysis, these lessons focus on epistemological, methodological and practical issues rather than answers to our central research questions. Nonetheless, we think that these lessons are worthy of being shared with colleagues since they are central to the challenges associated with any investigation of teaching and learning. We group these lessons under three broad headings: (1) unanticipated challenges, (2) the difficulty of capturing complexity without sacrificing clarity, and (3) the intrusiveness of the current policy environment.

Unanticipated Challenges

Our initial research design for the high quality teaching study acknowledges the complexity of teaching and the challenges posed to researchers interested in understanding how teacher practices relate to student learning. In retrospect, we didn't fully appreciate just how challenging studying teacher practices would be. Some of these challenges are generic, perhaps timeless, and thus have been experienced by other researchers who have attempted to understand what is teaching and how does teaching manifest itself in classrooms. Other challenges, though, may be more reflective of the current policy environment, a possibility that we return to later in the manuscript. Some of the more important challenges that emerged during the early implementation of the study, though, have to do with identifying "taken-for-granted" elements of teaching – namely, "who is the teacher," "what is the lesson," and "what is an instructional practice." As we began developing instruments and collecting data, we realized quickly that these three central elements of teaching are not easily identified.

Teacher. A fundamental premise of much of the current research on teaching is that

teaching quality is central to student learning. As a result, there is strong encouragement within the current policy environment to base individual teacher evaluations on contributions to student achievement gains. We, too, assumed a relatively straightforward connection between students and the teachers who provide them with instruction, but we soon came to ask, “*Just who is the teacher?*” A common conception at the elementary school level is that students are assigned to a classroom and receive their instruction from that classroom teacher. In some instances, especially in the upper elementary grades, there might be some departmentalization, with teachers who have particular expertise being assigned specific subject areas, such as science, mathematics, reading/language arts, or social studies. In either case, though, the assumption is that it is relatively easy to identify a student’s teacher in any subject area.

But in our participating schools we have found much more variation and complexity in how students and teachers are linked in the instructional process. First, even when teachers teach across all subject areas and are not departmentalized, they do not necessarily teach the same students. Students can be assigned to different teachers for different subjects based on prior achievement in those areas, desire for more homogeneous or heterogeneous groupings, attempts to reduce class size (especially for reading instruction), language proficiency, and so forth. Sometimes these assignments are made by the school’s administration; sometimes the grade level team makes them. Grade-level teachers sometimes pair up and switch students for particular units and then switch back again. During the course of a year, a student may be assigned, even if only temporarily, to two or three teachers.

Second, even when students stay in a single classroom for their instruction, someone other than the classroom teacher can teach them. Resource teachers, para-professionals, student teachers, staff developers, and team teachers all may contribute to a student’s learning (or mis-

understanding) of curricular content. We have observed in classrooms where the teacher of record consistently works with one reading group and para-professionals work with others, where a resource teacher or staff developer takes over part of the lesson to demonstrate a teaching strategy, and where the computer teacher pulls small groups of students to work on their writing assignment in the computer lab. Students can be in an assigned reading or math class for part of the time and sent to an ESOL or resource teacher for the rest of the time. Or the student can spend the entire instructional period with the classroom teacher and get a “double dip” of reading or mathematics during another part of the day with another teacher.

The question for researchers is whether these variations are just marginal noise, peripheral to the assigned teacher’s influence on student learning, or of such significance that they call into question basic research assumptions and measures of teacher effects. To begin to account for these multiple variations and influences we developed a web-based database for teachers to roster their students for classes in reading and mathematics; we use this system (and more conventional paper-and-pencil checklists) to take two “snapshots” of students assigned to teachers during the year. We also ask teachers to notify us of any changes in their rosters, to provide us with information about routine instructional assistance that they receive, and to identify individual students who receive regular resource help in or out of the classroom. This information will help us understand the network of influences on students’ learning but it will not resolve the methodological problem that many studies of teaching face – namely, that a substantial number of students have more than one teacher during the school year.³

³ Of course, this problem is well known with regards to student mobility between schools; what we are suggesting here is that there may also be substantially mobility within schools, an observation not addressed well in the current literature.

Lesson. Just as we have wondered about the link between students and teachers, we sometimes wonder, “*Just what is the lesson?*” It is common to think of lessons as discrete, time bounded segments of the day, in which teachers orchestrate students’ engagement with a specific subject for a set of instructional purposes. Of course, the more discrete and bounded the instructional event, the easier it is for observers to capture teacher and student engagement around a particular subject area, and it is easier for teachers to describe clearly and unambiguously curricular coverage as part of their daily logs. Although this model generally fits mathematics instruction in the classrooms that we have observed, where teachers engage students in mathematical knowledge for a continuous block of time (ranging from 60-90 minutes), it fits less well instruction in reading/language arts, where instruction occurs in varying blocks of time and is often blended into other subject areas.

Variations across schools and changes over time in reading/language arts instruction have made consistency in identifying and recording lessons more difficult than we imagined. In some of our schools, 90-120 minutes a day is set aside specifically for instruction in reading/language arts, though students and teachers deviate from this schedule roughly once a week to accommodate what are called “specials” (e.g., art, music, or physical education), to use the media center, or to coordinate access to the computer lab. In other schools, 60 minutes is set aside for reading in the morning and another 60 minutes for writing in the afternoon (or vice versa), and in still other schools, teachers routinely incorporate reading/language instruction into social studies or science instruction, essentially extending the “official” reading/language arts block into other parts of the day. Moreover, in some schools students engage in “Bell Work”, usually review of grammar, punctuation, spelling, and sentence structure but sometimes oral reading, while beginning-of-day routines take place.

The question our observers confront is where does a lesson “begin” and “end”? Does a “lesson” begin and end during an official block of time as orchestrated by an individual teacher, or is a “lesson” the accumulation of students’ engagement in a content area during the course of a given day. Our interview data with teachers, as well as our own observations, indicate that some teachers intentionally use “unofficial” blocks of time to “extend” lessons and reinforce foundational skills, especially in reading and writing. Other teachers extend lessons more serendipitously but with sufficient purpose to lead an observer to question whether the teacher has “ended” one lesson and “began” another. “Lessons” ebb and flow throughout the day in many of the schools that we have observed but without the regularity of ocean tides. So even though all our teachers and observers have been given the same assignment, to record the 60-90 minute reading and mathematics lessons presented to students in a particular class, we cannot be certain that we are always capturing the “lesson” – certainly not the whole lesson presented to students during the course of a day.

We have employed a range of strategies to address this ambiguity, particularly in our observations of reading/language arts lessons. First, when observers schedule observations, they privilege reading instruction in determining a time to observe teachers’ classes. Doing so provides some consistency across observations and increases the likelihood that we will capture “lessons” or “aspects of lessons” that tap a similar curricular focus. Second, we conduct pre-post interviews with teachers to determine what they perceive to be the boundaries for a particular lesson; this provides us with information about how a teacher conceived of a lesson and whether we have captured major aspects of its intended structure. Finally, we rely on teachers’ logs, which are typically a reflection of what a teacher did during the day in a curricular area to capture aspects of a lesson that we may not have had an opportunity to observe. Although these

strategies do not resolve the problem of “what is a lesson”, they increase the likelihood that we will capture major elements of daily instruction in a curricular area with our instruments.

Practice. Similar to the problem of defining a lesson, is the problem of defining the unique elements of instruction – that is, teacher practices. Although we think of practices as the interactions among teachers, students, curriculum and context directed toward some instructional purpose, such a definition still provides substantial room for ambiguity. From a researcher’s point of view, it is important to be able to identify what it is that teachers do with different curricular materials and in different contexts to facilitate student learning. But when does what teachers “do” in a classroom constitute “practice”? Does the use of whole-group instruction versus small-group instruction during a lesson constitute a specific practice or must the use of this particular form of instructional organization be seen in the context of how students and teachers interact around curricular content? Should we understand practice to be what teachers do at a specific point in time, what teachers do during the course of a lesson, or as what teachers do across multiple lessons during the course of a year?

Although the issue of what constitutes “practice” is endemic to all of our instruments, it became most apparent to us as we developed the Time Sampling protocol.⁴ Here the question is simply – what should we observe if we want to capture the essential elements of teacher practices? Recall that the Time Sampling protocol requires the observer to record what is happening in a classroom every 3 minutes during a lesson. We deliberately constructed a narrow temporal window for episodes, requiring observers to restrict their attention to a time period of no longer than about 15 seconds. By doing so we hoped to enhance interrater reliability and

⁴ We discuss this issue with regards to time sampling but researchers have raised similar concerns about other types of instruments, such as video taping of classroom instruction. Observers viewing the tape often arrive at different conclusions about the practices that characterize a teacher’s instruction, in part because the conceptual boundaries of practice are elastic and can be difficult to identify.

capture the basic elements of lessons and eventually practice. During an episode, observers often record what we refer to as “adjacent pairs” of action (or inaction) that take place immediately as the protocol screen opens, such as a teacher posing a particular type of problem and students responding in a particular way, or a student answering a question and the teacher requesting a particular type of response from the rest of the class. Along with information about student and teacher activity, observers record information helpful in understanding the content and context of instruction actions, classroom organization and the teacher’s attention, the use of materials and technology, and students’ behavior and engagement. Only that isolated 15-second slice of classroom life is recorded. Everything else, no matter how telling or provocative, is ignored.

We routinely discuss the capacity of these narrow slices of classroom life to reveal the complex act of teaching. But we believe the benefits of our decision to focus on specific episodes outweigh the potential losses. For example, the intent of an overall lesson could be to develop students’ understanding of a concept like elapsed time and to link their procedural knowledge to conceptual knowledge. Hearing the teacher discuss a lesson like that in advance predisposes you to think about a “linking” lesson. Without having to attend to exactly what the teacher and students are doing the moment the screen opens, it’s easy to miss how frequently a teacher might proceduralize the problem—break it down into small parts so students are following a pattern rather than figuring things out for themselves. By focusing on a series of relatively narrow time points within a lesson it is more difficult to impose patterns of practice that may not be there.

Furthermore, the thousands of episodes we record across time, lessons, and teachers, capture the deep structure of classroom life, not idiosyncratic occurrences, as well as prevalent

patterns of practice that occur within and across lessons and within and across subject areas. Within one lesson, we average 20 episodes. Six observations of a teacher give us 120 episodes in mathematics and 120 in reading. Seventy teachers observed six times give a sample total of 8,400 episodes a year in each subject area. Patterns and flow of classroom life are captured by these episodes when they are examined together. Teachers' implicit decisions with regard to how they engage students in curricular content within the context of the classroom can be seen in patterns of codings for lessons and content areas. Although we have just begun analyzing these data, we are confident that they will provide us with the means to identify patterns of practice associated with different teachers, curricular content, and classroom contexts.

Complexity and Clarity

Though not an unanticipated challenge, we have gained renewed appreciation for the difficulties involved in honoring the complexity of teaching without jeopardizing the clarity of research results. We have especially experienced this challenge in three areas of our study: (1) in our intention to examine not only generic but subject-specific consequences of teaching, (2) in our desire to maintain multiple perspectives on what constitutes quality, and (3) in our reliance on multiple conceptions and methodological approaches to studying teaching. Of these three areas the latter may be the most critical, since it captures the challenges associated with implementing a research design that acknowledges both the complexity of teaching *and* the desire for greater clarity about organizational and pedagogical practices that promote higher-levels of learning and greater educational equity in classrooms.

Generic v. subject-specific practices. As mentioned earlier, few studies directly compare the teaching of reading and mathematics. Most studies take either a generic perspective, ignoring the importance of subject matter in teaching, or focus exclusively on the teaching of one

subject or the other. But because elementary school teachers generally teach both mathematics and reading, comparative questions are of particular interest to practitioners and policymakers. One of the potentially unique contributions that the HQT study can make to our understanding of teaching is its careful attention to similarities and differences across two core subject areas at the elementary school level. We believe that this contribution can be realized both in the conceptualization and development of instruments meant to tap practices and in the empirical perspective that we will be able to provide on similarities and differences in reading and mathematics instruction that promote learning. However, even very preliminary analyses of our data indicate that interpreting similarities and differences in our measures requires considerable caution.

In developing data collection instruments, we realized that many important aspects of teaching mathematics and reading could be captured quite well by the same conceptual and measurement categories. In both the time sampling and teacher log protocols, items that characterize what teachers and students do, and how instruction is organized and contextualized, are often identical. These similarities suggest that scholarship in reading and mathematics instruction may provide similar conclusions – that teaching in reading and mathematics may actually “look” quite similar under certain circumstances. And that is, in fact, what preliminary analyses of our data indicate. In our first year of time sampling data, observers recorded similar profiles for student and teacher activities, regardless of subject area. Teachers also looked similar regarding the contextualization of teaching and classroom behaviors (Chambliss & Graeber, 2003). Although we have not examined the consequences of these practices for learning in the respective content areas, these results do indicate substantial overlap in how teachers approach instruction in reading and mathematics.

These similarities may, however, mask important differences in teacher practices and what teachers' understandings are of what they are attempting to accomplish in the classroom. As reported in Chambliss and Graeber (2003), preliminary data analyses reveal both similarities in what observers report and potentially intriguing differences in how the same actions by teachers relate to other items included in the Time Sampling protocol. Requesting that students elaborate on an answer in mathematics, for example, was highly correlated with responses indicating high-levels of cognitive engagement by students, while this same relationship did not hold for reading. Requesting an elaboration in mathematics may signal a different set of expectations in mathematics than in reading, requiring a greater depth of domain knowledge in mathematics than reading for both the teacher and the student providing the response. Further analyses of the larger data set will allow us to explore such differences more exactly.

The elusive concept of quality. Determining the quality of teaching has been a perplexing concern throughout the history of research on teaching, and not one that we are likely to abandon in the near future. Researchers invariably rely on proxy indicators: teacher characteristics, frequency measures, descriptive cases, and teacher perceptions. Most remote are the *teacher* characteristics: years of experience, education level, number of subject matter courses taken, type of teaching license and its match with teaching assignment. Other common indicators are various types of frequency measures that either correspond to commonly held standards of “good teaching” or that correlate with measures of student learning. These approaches, however, raise fundamental issues about how we assess “quality.” In other words, how *good*—of what quality—are the standards themselves? How good are the measures of student learning used to identify “quality” teachers? Even though we have constructed multiple ways of looking at quality teaching and its relation to student learning, these questions are ultimately answered on the

basis of human judgments, values, and limited knowledge.

Like many studies, the High Quality Teaching is heavily dependent on its measures, correlational analyses, and numerous external standards if it is to identify “quality” teacher practices. The Time Sampling observation protocols count the number of times teachers interacted with students in particular ways around particular types of subject matter. The Teacher Logs give frequency indicators of the enacted curriculum. Even the Attribution Scale, seemingly a high-inference measure of quality, has frequency, not quality, anchors: “not evident” and “pervasive.” Observers are told that the attributes are characteristics of lessons and that the nature of the lesson dictates whether or not the various attributes warrant a “pervasive” rating. We urge them not to equate a “1” (not evident) with low-quality teaching and not to simply average their ratings on the 20 items to arrive at the one overall quality rating. We also know that observers do not equally value all attributes. Some, for example, place a higher value on incorporating student interest and choice than others. For other observers, promoting principled understanding of subject matter is most important. These differences in observer values and beliefs undoubtedly influence their overall quality ratings.

In-depth case studies and teacher interviews are often considered to be a better way of getting at teaching quality, which is one reason why we included this methodology on our research design. In the High Quality Teaching study, we select classrooms for detailed case studies on the basis of multiple proxy indicators of quality: student learning gains, attribution ratings, and the recommendations of classroom observers. We also interview teachers for their perceptions of high-quality teaching and look for ways to incorporate these indicators into our study. Although we believe that the use of multiple indicators is better than privileging a single set of indicators of quality, these decisions ultimately raise the same types of question

about what constitutes “quality”. We have found no clear, easy way to study teaching quality without acknowledging multiple and even competing standards (not do we expect to find one). Our strategy has been to use multiple types of data and multiple standards but this approach creates other issues regarding the organization, analysis, and interpretation of data.

Multiple conceptions and measures of teaching. Because we have sought to capture the complexity of teaching and its variation across lessons, curricular areas, and classroom contexts, we have relied heavily on a research design that embraces different conceptions of teaching, different methods of studying teaching, and different epistemological standards that guide the investigation and interpretation of teaching as a phenomenon. There are substantial methodological strengths associated with the use of multiple instruments and methodological approaches. Multiple measures of teaching, for example, can be used to establish concurrent validity, while multiple methods of investigation can be used to capture different aspects, quite often different levels of detail, of a phenomenon. The key questions are not only how similar are interpretations of data that use different instruments but what do we learn that is different about a phenomenon by using a different method.

In the High Quality Teaching study, we have consciously developed instruments and relied on multiple methods to address both questions. For example, we have designed each instrument to capture distinct aspects or dimensions of teaching. In each case, our goal has been to make the instrument as similar as possible for both reading and mathematics while still honoring the unique nature of the two subjects. The daily teacher logs were designed primarily to gather information on curriculum coverage: depth, breadth, variety, and pacing. The time sampling instrument focuses primarily on instructional practices: the frequent, sequential interactions that students and teacher have around subject matter. In time sampling,

discrete narrow slices of teaching are recorded within the behavioral categories predetermined by researchers. The attribution scale, on the other hand, provides an overall gestalt of a reading or mathematics lesson, anchored explicitly in items that express five learner-centered learning domains.

We have also utilized multiple methods in the study, methods quite often identified as having distinct and possibly conflicting epistemologies. So to deepen our understanding of lessons, as well as to provide an alternative perspective to the time sampling protocols and attribution scales, we interview teachers before and after lessons to determine the meanings that teachers attribute to their actions and students' responses. We conduct individual and focus-group interviews with teachers to understand how they define quality teaching, and we interview principals to ascertain the broader policy context in which teachers make pedagogical decisions, adjust their practices, and influence learning outcomes for individual students and their classes. While these multiple methodological approaches to investigating teaching provide us with substantial detail and diverse perspectives on teaching, they also pose a problem – how do we integrate these perspectives, preserve the integrity of their epistemological positions, and enhance our understanding of teaching?

The challenges for such an approach are seeing patterns and commonalities across instruments, subject matter and contexts, while not sacrificing the unique insights that might be privileged by a particular set of data or a particular methodological approach. One approach that we have taken to this challenge is to link the various instruments that we have developed to a dominant theory of learning. The learner-centered principles, implicitly or explicitly, guide each of our data gathering and analysis strategies: time sampling, teachers' daily logs, attribution

scales, and case studies. A second strategy has been to engage researchers who are experts in different methodologies or responsible for different instruments in on-going discussions about data collection, analysis, and interpretation. Because of these underlying principles and ongoing discussions, we create the possibility for discovering common patterns of practice across the various instruments and methods, even though we expect the unique insights about practice from the different instruments and methodological perspectives on which we rely. We are optimistic about our ability to capture the complexity of teaching without jeopardizing the clarity of our findings; however, we do not underestimate the challenge associated with achieving the proper balance.

Press of High-Stakes Policy Environment

Studies of teaching are increasingly conducted in an environment characterized by high-stakes accountability. Teachers, principals, and schools are under *intense* pressure for students to do consistently better on standardized achievement tests and to close the persistent gap between higher and lower achieving groups. These pressures impact the work of researchers in many ways: teachers are apprehensive about being observed, concerned about the uses and abuses of data, protective of their time, and want immediate feedback for their efforts. Moreover, demands for accountability can result in dramatic changes in expectations for teachers and in the formats for assessing outcomes, creating practical and methodological issues about the validity of instruments and proposed analytic models.

Although we expected the policy environment to have an effect on the teachers in our study – indeed, capturing the effects of this environment on teachers’ efforts to scale up and sustain effective pedagogy is a central concern for the study – we did not anticipate the manner in which the policy environment would heighten and intensify some of the challenges associated

with any study of teaching, particularly a study of teaching over time. We highlight three challenges and responses: difficulties in recruiting and retaining participants, local policy changes that create practical and methodological problems, and challenges associated with maintaining the integrity of a study in the face of real needs by principals and teachers for practical solutions.

Recruiting and retaining participants. The longer a study of teaching, the greater the possibility that participants will drop out or move outside of the sampling parameters. While we anticipated some teacher change based on local district records, we were unprepared for the amount of turnover we experienced during the first three years of the study. In the first year, 11 schools and 67 teachers agreed to participate. Going into the second year we lost four of these schools and nearly half of the teachers. We refreshed the participant pool by adding 9 more schools and recruiting additional teachers in the second year; this increased the sample size to 16 schools and 73 teachers. But in the third year, we lost schools and teachers again. Once more we refreshed the pool, adding 5 new schools and recruiting 36 new teachers, bringing the total sample in the third year to 18 schools and 76 teachers. Of these teachers, however, only 12 have participated across all three years of the study, far fewer than we had anticipated.

What accounts for the attrition of schools and teachers? At the school level, factors included insufficient incentives to participate, changes in building leadership, fears that participation might detract from other priorities, and concerns about how the results of the study might reflect on the reputation of schools, a principal's leadership, and individual teachers. At the teacher level, we found three types of reasons for not participating or withdrawing: personal (e.g., health, maternity/family leave); a change in school or position that fell outside our sample parameters (e.g., assignment to a different grade-level or school district; or assignment to

a low-poverty school); and teachers feeling too overwhelmed by other responsibilities to participate. Although we've developed a series of strategies to address attrition, we anticipate additional losses in schools and teachers as we approach our final year of data collection.

Attrition poses many methodological and practical problems to research, not the least of which is the effort involved in “selling” a study to principals and teachers and then training them to participate effectively. We have developed new incentives to retain participants, established liaison positions to improve communications, promoted the use of e-mail to schedule observations, distributed newsletters to acknowledge the efforts of teachers, and created rapid-response strategies to respond more quickly to technological difficulties that teachers might experience maintaining daily logs. Although each of these strategies seems to have had a positive effect on participation, they do not resolve an essential problem associated with research on teaching in the current policy environment – principals and teachers are under mounting pressure to increase annually the performance of students in their schools and classrooms and participation in the High Quality Teaching study is at best a long term investment with unknown returns. Such an environment is not conducive to long-term commitments.

Local policy changes. In addition to the challenges associated with recruiting and retaining participants, our study has also faced changes in curriculum and assessment practices during the first three years. Each of these changes is a response to the mounting pressures of high-stakes accountability policies being implemented at the district, state, and federal levels. These changes are certainly of interest to us given our research questions; however, they also create practical and epistemological dilemmas in terms of adjusting our data collection plans and maintaining the validity of instruments. We were fortunate in being able to anticipate curriculum changes when we created our standardized protocols. If the changes had come a year later,

our time sampling and teacher log protocols might have been misaligned with the new curriculum. The real impact of the curriculum changes on the study came in the form of teacher stress and anxiety. The new mathematics curriculum was being created as teachers were asked to implement it. Sometimes teachers received new mathematics units just days ahead of the implementation schedule. This pressured timeline quite possibly had a negative impact on the quality of their teaching; it undoubtedly had an impact on their enthusiasm for participating in a study about high quality teaching.

Even more worrisome, from a research perspective, were the changes in standardized assessments that have occurred since we began the study. These changes involved both the types of assessments being used and the manner in which testing influences instructional opportunities. The original assessments we used to select schools were district-developed criteria reference tests (CRTs). We used these data to investigate average achievement gains given FARMS enrollments and annual changes in achievement between 1997 and 1999. After selecting the initial sample, however, MCPS decided to abandon the annual CRTs, relying on, instead, the state test (MSPAP) at the 3rd and 5th grade level and the CTBS at the 2nd and 4th grade level. As a result of the *No Child Left Behind* legislation, the state then discarded the MSPAP and, in 2002, implemented a new test, the Maryland State Assessment (MSA), given each year in grades 3-8. While the correlation of student achievement scores across these tests tends to be relatively high, about .70, the variation in achievement tests creates potential measurement error with unknown consequences. We are investigating the parameters for each assessment and developing statistical procedures for equating outcomes.

Threats to integrity. Perhaps one of the more important lessons that we have learned about the influence of the policy-environment on a study of teaching is how “high-stakes”

assessments ratchet up the pressure to respond to a real need on the part of principals and teachers for useful feedback. Although interactions between researchers and practitioners have always had some strain associated with differences in professional standards and motivation, a policy environment that requires dramatic improvements in student performance heightens these strains. Neither administrators nor teachers have much patience for generalized results or equivocal findings. They desire answers and assistance in meeting the demands for accountability. In this environment, a research dilemma is to present work with integrity: to resist pressure to report on work mid-stream, to protect confidences, and to guard against unintended interpretations or uses of data and research findings by policy makers.

We have increasingly come to appreciate the importance of finding ways of providing feedback without overstating or misleading practitioners about what we know and what we don't know regarding high quality teaching. We thought to diminish the tension between practitioner and research needs by making teachers "co-researchers" in our research design, but for many teachers this was interpreted less as an opportunity to examine their own practice and more as a burden—especially because we could not give them immediate feedback from the data collection. More successful have been feedback sessions, in which we discuss broad patterns in the log data collected by teachers, and annual debriefing meetings, in which we discuss with principal and teachers data collection activities, answer questions about the study, discuss broad patterns in the data, and express our appreciation for support. Nonetheless, even these events place an additional burden on time and fail to provide direct responses to principals and teachers questions about what they can do to enhance learning.

Of the challenges that we have experienced conducting the study of high quality teaching, this challenge may be the most foreboding and difficult to resolve. As

accountability requirements increase and intensify in the policy environment surrounding classrooms and schools, so does the gap between the interests of researchers and the interests of practitioners. Researchers require the cooperation and participation of practitioners if we are to understand what exactly constitutes high quality teaching and how it influences student learning in different subjects and contexts. Practitioners have a vested interest in this knowledge, especially if it is knowledge that can be generalized to the educational settings in which they work, but they require guidance in a timeframe and in a format that is not familiar to most researchers (e.g., professional development). The press of high-stakes accountability may find researchers looking into classrooms from afar, especially if principals and teachers see participation in studies as overly burdensome and disconnected from their immediate needs.

REFERENCES

- American Psychological Association Board of Educational Affairs. (1995, Dec.). *Learner-centered psychological principles: A framework for school redesign and reform*. Washington, DC: American Psychological Association. (<http://www.apa.org/ed/lcp.htm1>)
- Alexander, P. A., & Fives, H. (2000). Achieving expertise in teaching reading. In L. Baker, M. J. Dreher, & J. T. Guthrie (Eds.), Engaging your readers: Promoting achievement and motivation (pp. 285-308). New York: Guilford.
- Alexander, P. A., & Murphy, P. K. (1998). The research base for APA's learner-centered psychological principles. In N.M. Lambert & B.L. McCombs (Eds.), Issues in school reform: A sampler of psychological perspectives on learner-centered schools. Washington DC: The American Psychological Association.
- Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes, Teaching as the learning profession: Handbook of policy and practice (pp. 3-32). San Francisco: Jossey-Bass
- Bransford, J. D., Brown, A. L., Cocking, R. R. (Eds.) (1999). How people learn: Brain, mind, experience, and school. Washington DC: National Academy of Science.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 328-75). NY: Macmillan.
- Carpenter, T., Fennema, E., Peterson, P., Chiang, C., Loef, M. (1989). Using knowledge of children's thinking in classroom teaching: An experimental study. American Educational Research Journal, 29, 17-28.
- Chambliss, M., & Graeber, A. (2003, April). Does subject matter matter? Paper presented at the meeting of the American Educational Research Association, Chicago.
- Cobb, P., Wood, T., Yackel, D., Nicholls, J., Wheatley, F., Trigatti, B. & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. Journal for Research in Mathematics Education, 22, 3-29.
- Cohen, D. K., & Hill, H. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102, 294-343.
- Coleman, J. et al. (1966). Equality of educational opportunity. Washington DC: U. S. Government Printing Office.
- Croninger, R., Valli, L., & Price, J. (2003, April). Mapping the policy environment for high-quality teaching: Can we get there from here? Paper presented at the meeting of the

American Educational Research Association, Chicago.

Darling-Hammond, L. (1997). The right to learn: A blueprint for creating schools that work. San Francisco: Jossey-Bass.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo, J. (1999). NAEP 1998 reading report card for the nation and the states, executive summary. (<http://nces.ed.gov/nationsreportcard/pub/main1998/1999500.shtml>).

Dunkin, M., & Biddle, B. (1974). The study of teaching. NY: Holt, Rhinehart & Winston.

Fennema, E., Carpenter, T., Franke, M., Levi, L., Jacobs, V., & Empson, S. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. Journal for Research in Mathematics Education, 27, 403-434.

Ferguson, R. F. (1998) Can schools narrow the black-white test score gap? In C. Jencks & M. Phillips (Eds.), The black-white test score gap (pp. 318-374). Washington, D.C.: Brookings Institute Press.

Ferguson, R., & Ladd, H. (1996). How and why money matters: An analysis of Alabama schools. In Helen Ladd (ed.), Holding schools accountable (pp. 265-298). Washington D.C.: Brookings Institute Press.

Floden, R. (2001). Research on effects of teaching: A continuing model for research on teaching. In V. Richardson (Ed.), Handbook of research on teaching (4th ed., pp. 3-16). Washington DC: AERA.

Good, T., & Brophy, J. (2003). Looking in classrooms, 9th ed. Boston: Allyn & Bacon.

Greenwald, R., Hedges, L., & Laine, R. (1996). The effect of school resources on student achievement. Review of Educational Research, 66 (3), 361-396.

Hawley, W., & Rosenholtz, S. (1984). "Good schools: A synthesis of research on how schools influence student achievement." Special issue, Peabody Journal of Education, 4, 1- 178.

Hiebert, J. (Ed.) (1986). Conceptual and procedural knowledge: The case of mathematics. Hillsdale, NJ: Lawrence Erlbaum.

Hiebert, J., & LeFevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), Conceptual and procedural knowledge: The case of mathematics (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum.

Jackson, P. (1968). Life in classrooms. New York: Holt, Rinehart and Winston.

- Jencks, C. et al. (1972). Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books.
- Kilpatrick, J., Swafford, J., Findell, B.(Eds.) (2001). Adding it up: Helping children learn mathematics. Washington DC: National Research Council, National Academy Press.
- Kliebard, H. (1973). The question in teacher education. In D. McCarty & Associates. New perspectives on teacher education. San Francisco: Jossey-Bass.
- Lampert, M. (2001). Teaching problems and the problems of teaching. New Haven: Yale.
- McNeil, L. (2000). Contradictions of school reform: Educational costs of standardized testing. New York & London: Routledge.
- Morrow, L. M., Wamsley, G., Duhammel, K., & Fittipaldi, N. (2002). A case study of exemplary practice in fourth grade. In B. M. Taylor & P. D. Pearson (Eds.), Teaching reading: Effective schools, accomplished teachers (pp. 289-307). Mahwah, NJ: Erlbaum.
- National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.
- National Commission on Teaching and America's Future (1996). What matters most: Teaching for America's future. New York: Author.
- National Research Council. (1998). Preventing reading difficulties in your children. C. E. Snow, M. S. Burns, & P. Griffin (Eds.) Washington, DC: National Academy Press.
- Newmann, F.M., & Wehlage, G.G. (1995). Successful school restructuring. A report to the public and educators by the Center on organization and restructuring of schools. Madison, WI: University of Wisconsin-Madison.
- Porter, A. (2004, January). Measuring instructional alignment. Paper prepared for the conference on The Measurement of Instruction: Technical Challenges and Implications for Research, Policy, and Practice, Washington, DC: CPRE.
- Pressley, M., Rankin, J., & Yokoi, L. (1996). A survey of instructional practices of primary teachers nominated as effective in promoting literacy. Scientific Studies of Reading, 1, 145-160.
- Pressley, M., Yokoi, L., Rankin, J., Wharton-McDonald, R., & Mistretta, J. (1997). A survey of the instructional practices of grade 5 teachers nominated as effective in promoting literacy. Scientific Studies of Reading, 1, 145-160.

- Reese, C., Miller, K., Mazzeo, J., & Dossey, J. (Eds.). (1997). NAEP 1996 mathematics report card for the nation and the states: Findings from the National Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.
- Rowan, B. (2000, February). Assessing teacher quality: Insights from school effectiveness research. Paper prepared for meeting of the USDE Expert Panel: Strategies for Evaluating Efforts to Improve Preservice Teacher Education. Washington DC.
- Rowan, B. (2004). Using instructional logs to measure instruction in the study of instructional improvement. Paper prepared for the conference on The Measurement of Instruction: Technical Challenges and Implications for Research, Policy, and Practice, Washington, DC: CPRE.
- Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. Journal of Personnel Evaluation in Education, 12 (3), 247-256.
- Sarason, S. (1999). Teaching as a performing art. New York: Teachers College Press.
- Schoen, H., Hirsch, C., and Ziebarth, S. W. (1998). An emerging profile of the mathematical achievement of students in the Core-Plus Mathematics Project. Paper presented at the annual meeting of the American Educational Research Association, San Diego.
- Shafer, M., & Romberg, T. (1999). Assessment in classrooms that promote understanding. In E. Fennema & T. Romberg (Eds.), Mathematics classrooms that promote understanding (pp.185-200). Mahwah, NJ: Erlbaum.
- Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15 (2), 4-14.
- Skemp, R. (1987). The psychology of learning mathematics (expanded American Edition). Hillsdale, NJ: Erlbaum.
- Stipek, D., Salmon, J., Givven, K., Kazemi, E., Saxe, G., & MacGyvers, V. (1998). The value (and convergence) of practices suggested by motivation research and promoted by mathematics education reformers. Journal for Research in Mathematics Education, 29 (4), 465-488.
- Taylor, B. M., Pearson, P. D., Clark, K. F., & Walpole, S. (1999, September 30). Beating the odds in teaching all children to read: CIERA Report #2-006. Ann Arbor: Center for the Improvement of Early Reading Achievement, University of Michigan.
- Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2002). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-

income schools. In B. M. Taylor & P. D. Pearson (Eds.), Teaching reading: Effective schools, accomplished teachers (pp. 3-72). Mahwah, NJ: Erlbaum.

Valli, L. (2000). "Facilitating Reading Instruction Through School-Wide Coordination." In L. Baker, J. Dreher, & J. Guthrie (Eds.) Engaging young readers: Promoting achievement and motivation (pp. 237-263). New York: Guilford Press.