Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement

Heather C. Hill
Brian Rowan
Deborah Loewenberg Ball

Paper presented at the 2004 annual meeting of the American Educational Research Association, San Diego, CA; April 12, 2004. The authors would like to thank Robert J. Miller, Geoffrey Phelps, Stephen G. Schilling and Kathy Welch for their assistance. Errors remain the property of the authors.

The research reported in this paper was supported in part by grants from the U.S. Department of Education to the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania (Grant # OERI-R308A60003) and the Center for the Study of Teaching and Policy at the University of Washington (Grant # OERI-R308B70003); the National Science Foundation's Interagency Educational Research Initiative to the University of Michigan (Grant #s REC-9979863 & REC-0129421), The William and Flora Hewlett Foundation and The Atlantic Philanthropies. Opinions expressed in this paper are those of the authors, and do not reflect the views of the U.S. Department of Education, the National Science Foundation, the William and Flora Hewlett Foundation or the Atlantic Philanthropies.

Abstract

This study explored whether and how teachers' mathematical knowledge for teaching contributes to gains in student achievement. To do so, we used a three-level hierarchical model in which first (n=1190) and third (n=1773) graders' mathematical gains over a year are nested within teachers (n=334 and n=365), who in turn are nested within schools (n=115). We found that teachers' mathematical knowledge for teaching significantly contributed to student gains in both grades, controlling for key student and teacher-level covariates such as student SES, student absence rate, teacher credentials, teacher experience, and average length of mathematics lesson. While this result is consonant with findings of the educational production function literature, our measure of teacher mathematical knowledge is different from the measures of basic computational ability typically used in that literature. Instead, our result is obtained with measures of the specialized mathematical knowledge and skills needed in teaching mathematics. This finding provides support for policy initiatives focused on improving teachers' specialized knowledge for teaching mathematics.

In recent years, teachers' knowledge of the subject matter they teach has attracted increasing attention from policymakers. To provide students with "highly qualified teachers," No Child Left Behind requires teachers to demonstrate subject-matter competency through subject matter majors, certification, or other means. Programs such as California's Professional Development Institutes and the National Science Foundation's Math-Science Partnerships are aimed at providing content-focused professional development intended to improve teachers' content knowledge. This focus on subject matter knowledge has arisen, we argue, because of evidence suggesting U.S. teachers lack essential knowledge for teaching topics in mathematics (e.g., Ball 1990; Ma 1999), and because of evidence from the educational production function literature that teachers' intellectual resources significantly affect student learning.

Despite this widespread interest and concern, what counts as "subject matter knowledge for teaching" and how it relates to student achievement has remained inadequately specified. A closer look at the educational production function literature reveals that most studies operationalize teachers' knowledge with either proxy variables, such as courses taken or degrees attained, or tests of basic skills. Other scholars, however, conceptualize teachers' effects on students differently, arguing that it is how teachers know subject matter in light of the special tasks of teaching that drive any effects on student achievement (Ball 1990; Shulman, 1986; Shulman, Richert & Wilson, 1987). This specialized knowledge goes beyond that captured in measures of courses taken or basic skills. In mathematics, for instance, teachers may not only need to calculate correctly, but also need to represent numbers and operations using pictures or diagrams, provide explanations for common rules and procedures, and analyze students' mathematical solutions and explanations. By inadequately measuring teacher knowledge, the existing educational production function literature has been limited in the conclusions it might draw how teacher knowledge matters.

Further, in only a few of the educational production function studies (Harbison & Hanushek, 1992; Mullens, Murnane & Willett, 1996; Rowan, Chiang & Miller, 1997) was teachers' mathematical knowledge a predictor of student's mathematical achievement – most others use tests of teacher verbal ability to predict students' verbal and/or mathematical achievement. In fact, despite conventional wisdom that elementary U.S. teachers' subject matter knowledge influences student achievement, no large-scale studies have demonstrated this empirically (Wayne & Youngs, 2003). The situation is not made better by examining the educational process-product literature, where both the measurement of subject-specific teacher behaviors and the direct measurement of teacher knowledge were absent.

This study analyzes the relationship between teachers' scores on a measure of mathematical knowledge for teaching and their students' achievement. By "mathematical knowledge for teaching," we mean the mathematical content and skills that enable individuals to carry out the work of teaching mathematics. Examples of this work include explaining terms and concepts, interpreting students' statements and solutions, judging and correcting textbook treatments of particular topics, using representations accurately, and providing examples. A multiple-choice measure composed of items representing these teaching-specific mathematical skills has been shown to both reliably discriminate among teachers and meet basic validity requirements. Here, we model student gains in mathematics achievement as a function of teacher knowledge as operationalized by this new measure, controlling for student, teacher and school characteristics.

Framing the Problem

Since the 1960s, scholars and policymakers actively have explored the relationship between teacher characteristics, behaviors, and student achievement. Yet measures of teacher characteristics have varied widely, as have results from these investigations. Below, we outline how different literatures have measured teacher and teaching characteristics, and briefly summarize results from such investigations.

Teachers in the Process-Product Literature

In many ways, attempts to predict student achievement began with what is called the process-product literature, or the set of studies describing the relationship between teacher behaviors and student achievement. Moving beyond using affective factors such as teacher appearance and enthusiasm to predict student achievement, scholars in this tradition took the view that what teachers did in their classrooms might affect student achievement. By the late 1970s, scholars had accumulated substantial evidence that certain behaviors did matter. Focusing class time on active academic instruction rather than classroom management, student choice/game time, personal adjustment, or non-academic subjects was one quite clear correlate of student achievement; so was presenting materials in a structured format via advance organizers, making salient linkages explicit, and calling attention to main ideas. Brophy & Good (1986), Gage (1978), Doyle (1977) and others provide excellent reviews of these studies and findings. As this research progressed, scholars also designed experiments, training teachers in the principles and behaviors indicated by previous research and comparing their students' performance to those of untrained teachers. Notably, Good, Grouws, & Ebmeier (1983) performed such an experiment in mathematics and found that teachers who employed best practices had students who performed better in basic skills but not problem-solving.

Critiques of these studies ranged from methodological – e.g., relying on the use of correlational studies – to conceptual. Chief among the conceptual critiques included a lack of attention to subject matter, and how subject matter conditions the findings described above; what works well in mathematics, for instance, may not work well work in reading. Critics also pointed to the lack of subject matter knowledge in another aspect of this work, that is, scholars' failure to pay attention to teachers' subject matter knowledge as a predictor of effective teaching and learning. We discuss below how this critique – and this literature more generally – shaped our study and its findings.

Teachers in the Educational Production Function Literature

At the same time process-product studies were examining classroom behaviors and student achievement, other studies began to focus on the relationship between educational resources and outcomes. These studies, collectively called the "educational production function" literature, predicted student achievement on standardized tests from resources held by students, teachers, schools and others. Key resources include student background and educational histories, district financial commitments to teacher salary and material resources, and teacher and classroom characteristics (Hanushek, 1981; Greenwald, Hedges & Laine, 1996). Studies focusing on teacher characteristics employed two approaches, sometimes in combination, to measure the resources teachers bring to the classroom. In the first approach, information about teacher preparation and experience is collected and used as a predictor of student achievement. Key measures include teacher education level, certification status, number of post-secondary subject matter courses taken, number of teaching methods courses taken, and years of experience in classrooms. By using such measures, researchers implicitly assume a connection between formal schooling and experience and the more proximate aspects of teachers' knowledge and performance that lead to student outcomes. Reviews of this work dispute whether and how teacher preparation and experience contribute to student achievement (Begle 1972, 1979; Greenwald, Hedges, & Laine, 1996; Hanushek, 1981; 1996), with conflicting interpretations resting on the sampling of studies and methods for conducting meta-analyses. One potential reason for these conflicting findings, however, is that preparation and experience are poor proxies for the teacher resources and classroom behaviors that matter for helping students learn content.

A smaller number of educational production function studies employ a second approach, that of measuring teachers' resources by their performance on certification exams or other tests of subject matter competence. By using such measures, these studies implicitly assume a relationship between teacher content knowledge and classroom performance that lead to student outcomes. Studies using this approach tend to find a positive effect of teacher content knowledge on student achievement (e.g., Boardman, Davis & Sanday, 1977; Ferguson 1991; Hanushek 1972; Harbison & Hanushek, 1992; Mullens, Murnane & Willett, 1996; Rowan, Chiang & Miller, 1997; Strauss & Sawyer, 1986; for an exception, see Summers & Wolfe, 1977; for reviews, see Greenwald, Hedges & Laine, 1996; Hanushek, 1986; Wayne & Youngs, 2003).

Although the linking of teacher content knowledge with student achievement is an important research finding, this literature cannot fully describe how teacher knowledge relates to student achievement. One reason is that the literature offers only limited subject matter coverage: while the relationship of teachers' verbal ability to student achievement has been documented by many studies, only three studies have focused explicitly on both teachers' and students' mathematical knowledge and achievement (Harbison & Hanushek, 1992; Mullens, Murnane & Willett, 1996; Rowan, Chiang & Miller, 1997). A second reason is the design of many of these studies limits their potential for generalization. Two of the mathematics studies cited above take advantage of an assumed greater variation in teacher preparation and ability in other countries to identify content knowledge effects on student status or change (Harbison & Hanushek, 1992; Mullens, Murnane & Willett, 1996). Although these analyses are fundamental to building the theoretical case for the importance of teacher knowledge in producing student achievement, how these studies translate to U.S. contexts, where teacher preparation and knowledge may be both higher and more uniform, remains unknown. Other studies aggregate data to the school or district level, many analyze only cross-sectional data, and still others use composite measures of teacher knowledge or student achievement. Such methodological flaws limit the generalizations that can be made from such studies.

From our perspective, however, the most pressing problem in these studies remains the definition and measurement of teachers' intellectual resources, and by extension, the mis-specification of the models involved. Measuring quality teachers through performance on tests of basic verbal or mathematics ability may overlook key elements in what produces quality teaching. Effectiveness in teaching resides not simply in the knowledge a teacher holds personally but how this knowledge is used in classrooms. Teachers highly proficient in mathematics or writing will only help others learn mathematics or writing if able to use their own knowledge to perform the tasks they must enact as teachers —– for example, to hear students, to select and make use of good assignments, and to manage discussions of important ideas and useful work on skills. Yet these additional content-related abilities specific to the work of teaching have not been measured or included in the educational production function models. Harbison and Hanushek (1992), for instance, administered the same 4th grade math assessment to teachers and students, using scores from the first group to predict performance among the second. Mullens, Murnane, and Willett (1996) used teachers' score recorded on the Belize National Selection Exam, a primary-school leaving exam[1] administered to all students seeking access to secondary school. Rowan, Chiang and Miller (1997) used a one-item assessment of teacher knowledge; however, because no scaling or validation work was done on that item, little can be said about what and how well it measures. While the results of all three studies strongly indicate the importance of teachers' knowledge, we argue that recent theoretical work in how teachers' content knowledge matters for the quality of teaching leads to a need for measures more closely attuned to the mathematical knowledge used in teaching. We turn next to this literature in order to elaborate our argument.

Teachers in the Teacher Knowledge Literature

---

[1] The BNSE measures student proficiency at age 14, the equivalent in the U.S. of an end-of-eighth-grade exam.

This literature, focused directly on teacher knowledge, asks what teachers need to know about content in order to teach it to students. Researchers propose to distinguish between the ways in which teachers need to know content from the ways in which ordinary adults know such content. Shulman (1986; 1987), and his colleagues (e.g., Wilson, Shulman, & Richert, 1987) animated this line of inquiry with their groundbreaking work on what accomplished teachers know. In his 1986 presidential address to the American Educational Research Association, Shulman originally proposed three categories of teacher subject matter knowledge. His first category, <u>content knowledge</u>, was intended to denote "the amount and organization of knowledge . . . in the mind of teachers" (p.9). Content knowledge, according to Shulman, included both facts and concepts in a domain, but also why facts and concepts are true, and how knowledge is generated and structured in the discipline (Bruner, 1960; Schwab, 1961/1974). The second category advanced by Shulman and his colleagues (Shulman, 1986; Wilson, Shulman, & Richert, 1987) was <u>pedagogical content knowledge</u>. With this category, he went "beyond knowledge of subject matter per se to the dimension of subject matter knowledge <u>for teaching</u>" (p. 9, emphasis added). The concept of pedagogical content knowledge attracted the attention and interest of researchers and teacher educators alike. Included in pedagogical content knowledge, according to Shulman (1986), are representations of specific content ideas, as well as an understanding of what makes the learning of a specific topic difficult or easy for students. Shulman's third category, <u>curriculum knowledge</u>, involves awareness of how topics are arranged both within a school year and over time and ways of using curriculum resources, such as textbooks, to organize a program of study for students.

Shulman and colleagues' work expanded ideas about how knowledge might matter to teaching, suggesting that it is not only knowledge of content but also knowledge of how to teach such content that conditions teachers' effectiveness. Working in depth within different subject areas — history, science, English, mathematics — scholars probed the nature of the content knowledge needed by teachers. In this program of work, comparisons across fields were also generative. Grossman (1990), for example, articulated how teachers' orientations to literature shaped the ways in which they approached texts with their students. And Wilson and Wineburg (1988) showed how social studies teachers' disciplinary backgrounds — political science, anthropology, sociology — shaped the ways in which they represented historical knowledge for high school students. In mathematics, scholars showed that what teachers would need to understand about fractions, place value, or slope, for instance, would be substantially different from what would suffice for other adults (Ball, 1988, 1990, 1991; Borko, Eisenhart, et al., 1992; Leinhardt & Smith, 1985).

Until now, however, it has not been possible to link teachers' professional knowledge, conceptualized in these more subtle ways, to student achievement. Most of the foundational work on teacher knowledge described above has relied principally on teacher case studies (e.g., Grossman, 1990), expert-novice comparisons (Leinhardt & Smith, 1985), international comparisons (Ma, 1999), and studies of new teachers (Ball, 1990; Borko et al., 1992). Although such methods have been essential in beginning to specify the mathematical content knowledge needed by teachers, they lack the power to propose and test hypotheses regarding how elements of such knowledge contribute to helping students learn. The result has meant that although many assume, based on the educational production function literature, that teachers' knowledge does matter in producing student achievement, what and how knowledge matters is not well defined. Although many in the business of designing both professional development and policy assume that teachers' pedagogical content knowledge matters, exactly what this knowledge is, and whether and how it affects student learning has not yet been empirically established.[2]

To address these issues, the Study of Instructional Improvement (SII) began in 1999 to design measures of K-6 teachers' knowledge for teaching mathematics. In response to the above-reviewed literature, we focused our efforts on producing an instrument that could measure the knowledge used in teaching elementary school mathematics (Ball & Bass 2000, 2003). By "used in teaching," we meant to capture not only the actual content teachers taught – ordering decimals, or long division – but also the specialized knowledge needed for work with students. Specialized knowledge might include, for instance, how to represent quantities such as 1/4 or .65 using diagrams, or evaluate multiple solution methods for a problem such as 35 x 25. The desire to design measures of knowledge used in teaching also led us to construct items centered on the content of the K-6 curriculum, and the mathematical issues that arise in the course of teaching that content (e.g., providing a mathematically careful explanation of divisibility rules, or choosing an appropriate definition of "rectangle"), rather than items that might appear on a middle- or high-school exam for students. Details on measure design, construction, and scaling are presented below.

---

[2] We note, too, that "pedagogical content knowledge" itself has yet to be precisely defined and mapped. See Ball, (1988), Shulman (1986), Shulman (1987), Grossman (1990) and Wilson (1988) for different potential organizations of this knowledge.

Method

In this section we provide an overview of this project, describe the sample of students and teachers participating in the study, and detail data collection instruments and response rates. We offer more background on key teacher measures – content knowledge for teaching, teacher preparation, and teacher experience. Last, we explain data analysis methods and specifications, including why we chose to focus this analysis narrowly on teacher and student characteristics, rather than also include content covered, instructional practices and the school reform programs many sample schools were participating in.

Sample

The data explored here result from a study of schools engaged in instructional improvement initiatives. To enact this study, researchers collected survey and student assessment data from students and teachers in 115 elementary schools from the 2000-01 through 2003-04 school years. Eighty-nine of these schools are participating in one of three leading school improvement programs— America's Choice, Success for All, and Accelerated Schools Program—with roughly thirty schools in each program. Additionally, 26 schools not participating in one of these programs serve as comparison schools. Program schools were selected for the study via probability sample from lists supplied by the parent programs,[3] with some geographical clustering to concentrate field staff resources. Comparison schools were selected to match program schools in terms of community disadvantage and district setting. Once schools agreed to participate in this study, project staff approached all classroom teachers in each school to encourage their involvement.

---

[3] The sampling technique used conditioned school selection on geographic location, year of entry into CSR program, and an index of community disadvantage. The last ensured comparable schools within each CSR program. For more detail on the sampling process, see Benson (2002).

While the sampling procedure focused on schools engaged in instructional improvement, our achieved sample of schools and students only slightly over-represents high-poverty elementary schools and students and well represents the variability in the population of schools and students nationally. Where 1999 statistics show the average U.S. school serves an area where 13% of the households are in poverty, the average project school serves an area where 19% of the households are in poverty (Benson, 2002). Project schools are, however, disproportionately located in urban areas (47%) or mid-sized cities (21%). Comparisons of kindergartners enrolled in the first year of SII and in the Early Childhood Longitudinal Study (ECLS) show similarities in total family income and parental education level, but moderate disparities in student race and family structure (see Table 1). ECLS is a nationally representative sample, and thus this comparison suggests that our Year 1 sample did not enroll an entirely unique population of students. Importantly, sufficient variability in students and schools exists to suggest we are not working with a truncated sample, but instead one that over-represents non-white students residing in non-intact families in slightly higher-poverty areas. This suggests our sample represents a sufficient range of children, schools, and educational contexts to make reasonably confident statements about the contribution of teachers' knowledge to student achievement. Total, there were 1,190 students with complete data in the first grade cohort, and 1,773 students with complete data in the third grade cohort.

Project schools were located in 42 districts in 15 states. States varied in size, state average NAEP score, and in their approaches to improving low-performing schools. While 3 states scored as among the least interventionist on the accountability index designed by Carnoy and Loeb (2002), another 4 states scored at the top of this scale, indicating they were pursuing strong state-level rewards and sanctions to improve schools and student performance. The remaining 8 clustered near the less interventionist end of the Carnoy and Loeb scale. In one state and several districts, participation in a comprehensive school reform was mandatory for schools performing below a certain level; in other states and districts, comprehensive school reforms were entirely optional.

Our teacher sample for this paper included 334 and 365 first and third grade teachers, respectively. These teachers are fairly typical of the elementary teaching force, particularly in urban schools. On the year 2 questionnaire, (86%) of teachers reported they were female; a modest majority (55%) was White, with another 23% of teachers Black and 9% Hispanic.

Data Collection Instruments and Measures

At the center of the data collection effort are two cohorts of students. One cohort entered the study as kindergarteners and will exit the study at the end of second grade. A second cohort entered the study as third graders and will exit the study at the end of fifth grade. For each cohort, data were collected in two waves; in 2000-2001, the study collected information on Wave 1 first and third graders in 53 schools. In 2001-2002, the study collected information on Wave 2 first and third graders in the remainder of participating schools. These two waves have been collapsed in the data analyses, yet require the reporting of two years of response rates for most instruments (see below).

Before beginning descriptions of key variables, we briefly outline major instruments from which many of these variables derive, and provide information about response rates to these instruments. Information about students came from two major sources: student assessments and parent interviews. Student assessments were administered in the fall and spring of every academic year, for a maximum of six administrations over the life of the study. These assessments were given to eight randomly selected students per classroom, and administered outside students' usual classroom by trained project staff. Project staff also contacted parents/guardians of these target students once by telephone to gather information about students' academic history, parent employment status, and other relevant variables. The completion rate for the student assessment (Terra Nova) averaged 96% in the 2000-2001 and 2001-2002 school years.[4] The completion rate for the parent interview portion of the study was 85% and 76% in 2000-2001 and 2001-2002, respectively.

Teacher participation in this study included several components, most notably logging mathematics instruction up to 60 times during one academic year. The log is a daily self-report instrument that asks teachers to record the amount of time devoted to mathematics, the content topics covered, and the instructional practices used to teach mathematics. Teachers filled out logs for six-week periods in the fall, winter, and spring. Each log recorded one day of learning opportunities provided to one of eight randomly selected target students, the same students for whom SII also collects longitudinal achievement data. For log data, response rates are quite high; 97% (2000-2001) and 91% (2001-2002) of eligible teachers participated in efforts to record students' mathematics instruction through a daily log, and of the roughly 60 logs assigned per teacher, 91% were completed and returned to project staff.

The mathematics log used here has been subject to extensive development, piloting, and validation work. An observational study of an earlier piloted log suggested

---

[4] Fifty-three studies began student testing and teacher logging in 2000-2001; the rest began the study in 2001-2002, yielding two "cohorts" of students. We are grateful to the schools, teachers, and students participating in this study for allowing the collection of this data.

that agreement rates between teachers and trained observers was 79% for large content descriptors (e.g., number, operations, geometry), 73% for finer descriptors of instructional practice (e.g., instruction on why a standard procedure works), and that observer and teacher reports of time in mathematics instruction differed by under 10 minutes of instruction for 79% of lessons (Ball, Camburn, Correnti, Phelps, & Wallace, 1999).

Teachers were also asked by project staff to complete annual questionnaires about their background, involvement in and perceptions of school improvement efforts, professional development, and language arts and mathematics teaching. This survey contained the items included in the content knowledge for teaching mathematics measure described below. Table 2 also shows that roughly three-quarters of eligible teachers returned completed teacher questionnaires each year; because most of the non-content knowledge questions (e.g., on certification) remained the same on each teacher questionnaire, we were able to construct many of the variables described below even in the absence of one or two years of data.

Having described major instruments and response rates, we next turn to the specific measures used in this analysis. We begin with student achievement measures and working outward to those capturing characteristics of families, teachers, classrooms, and schools. Table 3 shows means and standard deviations for the measures discussed below.

Student achievement. The measure of student achievement used here was CTB/McGraw Hill's Terra Nova Complete Battery (spring of kindergarten), Basic Battery (spring of 1st grade), and Survey (third and fourth grade). Students were assessed in both the fall and spring of each grade by project staff members, and student scores are computed by CTBS using item response theory (IRT) scaling procedures. These scaling procedures yield interval-level (linear) scores from student's raw responses. For the analysis described here, we computed gain scores for the first year of study participation. For the first grade cohort, we subtracted each student's spring of kindergarten mathematics score from their spring of first grade mathematics score. For the third grade cohort, for which spring 2nd grade data was not available, we subtracted the fall of third grade mathematics score from the fall of fourth grade score. The result in both cases was a number representing how many scale score points students gained over one year of instruction.

The Terra Nova is widely used in state and local accountability and information systems. Its use here, therefore, adds to the generalizability of study results in the current policy environment. However, the construction of the Terra Nova adds several complexities to this analysis. To start, data from the mathematics logs indicate that the average student has a 70% chance of encountering number concepts, operations, or pre-algebra and algebra in any given lesson (Rowan, Harrison & Hayes 2004). For this and other reasons, our mathematics knowledge for teaching measure is constructed solely of items on these three focal topics. However, the Terra Nova samples mathematical topics more evenly across the elementary curriculum. At level 10 (spring kindergarten), 43% of items cover the focal topics. At level 12 (fall third grade), 54% of items cover the focal topics. As this implies, there is imperfect alignment between our measures of mathematical knowledge for teaching and measures of students' mathematical knowledge. Imperfect alignment between measures can cause underestimates of effect sizes, and suggests our models may underemphasize the importance of teachers' content knowledge in producing student gains.

Student mobility also affects the shape of this dataset, and can be examined from three different perspectives. By design, SII collects student achievement data on eight randomly selected students per classroom. Students who leave the classroom are replaced through random selection, but neither the leavers nor new students have complete data across the time points included in this analysis. Student mobility results in the completion of data for only 3.9 students per classroom in the first grade cohort, largely because mobility is always largest from K to 1, and 6.6 students per classroom in the third grade. Available information on student attrition helps inform the first perspective on attrition, namely that of asking who left: first graders who left the study scored 7 points lower on the spring kindergarten Terra Nova as compared to those with complete data across both time points; for third graders, the corresponding difference was 6 points. This difference in pre-test score was significant in logistic regressions predicting student attrition. Comparisons also showed that African-American and Asian students left the study at above-average rates. However, student attrition was unrelated to teacher scores on the content knowledge assessment in the third grade ($t=.282$, $p > .5$); that is, students who left the study were no more likely to have low-performing teachers than those who remained in the study. This relationship cannot be calculated for the first grade cohort, as the initial student mathematics assessment took place in spring of kindergarten, while students were with teachers for whom we have no content knowledge data. The lack of relationship at the third grade, however, suggests student attrition may not pose a major threat to the validity of findings.

A second concern when missing data occurs is whether the standard deviations of key independent variables are affected by the loss of portions of the student population; if this is the case, standardized regression coefficients could be biased. A comparison of key student-level variables (SES, minority status, gender, initial test score) using pre- and post-attrition samples shows standard deviations vary less than 5% in the case of initial test score, and only by 1% or less for the other variables. Because only .5% of first grade teachers 4% of third grade teachers had no complete student data, the standard deviations of teacher-level variables likely remain very close in both pre- and post-attrition samples.

Finally, although attrition was more common among students who performed more poorly on the initial pre-test, there remain in the data similar students to those who left the study. In this case, the growth of the "class" of lower-performing students can be accurately estimated, particularly given the independence of the probability of attrition and teachers' content knowledge, the main concern of this paper. All three lines of reasoning above suggest that student attrition does not pose a threat to the validity of results reported here.

Student background. The rate of student absence from mathematics instruction was generated by aggregating log reports of daily student absence to the student level. Just over 9 logs were recorded for the average first grader and 8.0 logs for the average third grader, and the reliability of this aggregated estimate in discriminating among students' rate of absence is .41. We then created a dummy variable marking students whose absence rate exceeded 20%. Information on students' gender/minority status was collected from teachers and other school personnel at the time of sampling. Information on family socio-economic status was collected via the telephone interview with the parent/legal guardian of the students in our study. The composite variable SES represents an average of father's education level, mother's education level, father's occupation, mother's occupation, and family income.

Teacher background and classroom characteristics. Teacher background variables came primarily from the teacher questionnaire, where data were used to construct measures of teacher experience, certification, and undergraduate/graduate coursework. Teacher background characteristics were straightforwardly represented in our models. For instance, teachers' experience was reported as the years in service during year 2 of the study. Although we had information on non-certified teachers' credentials (e.g., provisional or emergency certification), too few teachers existed in each category to include them independently in statistical analyses; thus our credential variable simply reports the presence (1) or absence (0) of certification. Finally, teachers reported the total number of a) mathematics methods and b) mathematics content courses taken on the questionnaire. However, since reports of methods and content courses correlated highly (r = .80) they produced multicolinearity in regression models estimated at both the first and third grade. As a result, we formed a single measure combining reports of mathematics methods and content coursework. Unfortunately, this strategy does not allow the examination of our models for the independent effects of methods and content courses, as is standard practice in the educational production function literature.

We included three classroom-level variables in this analysis. First, information on classroom percent minority students was obtained by aggregating student characteristics for each classroom. Second, to capture variation in the absolute amount of mathematics instruction students are exposed to, this analysis uses a measure of average time spent on mathematics derived from teachers' mathematics logs. The time measure excludes days on which the student or teacher was absent when representing the average minutes of mathematics instruction per classroom over the school year. Finally, the rate of teacher absence from mathematics lessons was also recovered by aggregating logs to the teacher level.

Content knowledge for teaching. Between five and twelve items designed to measure teachers' knowledge for teaching mathematics (CKT-M) were included on each teacher questionnaire. As the small number of items included each year suggests, the strategy was not to construct a reliable yearly measure of mathematical knowledge for teaching, but to construct one overall measure using data from multiple teacher questionnaires. Key to measure construction, however, was our desire to represent the knowledge teachers use in classrooms, rather than general mathematical knowledge, by designing tasks that gauged proficiency at providing students mathematical explanations, representations, and working with unusual solution methods. A more detailed description of the work of designing, building, and piloting these measures can be found in (Hill, Schilling & Ball, (2004). However, aspects of these measures are critical to interpreting the results of our analysis, and we highlight several features below.

Our work toward measuring content knowledge for teaching began with a two-step specification of a domain map. As noted earlier, we limited item-writing to only the three most-often taught mathematical content areas: number, operations, and patterns, functions, and algebra. Next, we decided what aspects of teachers' pedagogical content knowledge to measure within these three topics. Based on a review of the research literature, we chose to write items in only two major headings within the original PCK framework: knowledge of mathematics for teaching, and knowledge of students and mathematics.

Once the domain map had been specified, we invited mathematics educators, mathematicians, professional developers, project staff and former teachers to write items. Writers cast items in multiple-choice format to facilitate the scoring and scaling of large numbers of teacher responses, and produced items that were not ideologically oriented – eschewing, for example, items where a "right" answer indicated an orientation to "reform teaching." Finally, writers strove to capture two key elements in mathematical knowledge for teaching – teachers' "common" knowledge of content, or simply the knowledge of the subject a proficient student, banker, or mathematicians would have; and knowledge that is "specialized" to teaching students mathematics.

Two sample items illustrate this distinction (Figure 1). In the first, respondents are asked to determine the value of x in $10^X=1$. This is mathematics knowledge teachers use; students learn about exponential notation in the late elementary grades, and teachers must have adequate knowledge to provide instruction on this topic. However, many adults, and certainly all mathematicians would know enough to answer this item correctly – it is "common" content knowledge, not specialized for the work of teaching. Consider, however, another type of item. Here teachers inspect three different approaches to solving a multi-digit multiplication problem – 35 x 25 – and assess whether those approaches would work with any two whole numbers. To respond to this situation, teachers must draw on mathematical knowledge – inspecting the steps shown in each example to determine what the student has done, then gauging whether the method makes sense and works for all whole numbers. Appraising nonstandard solution methods is not a common task for adults who do not teach. Yet this task is entirely mathematical — not pedagogical; in order to make sound pedagogical decisions, teachers must be able to size up and evaluate the mathematics of these alternatives — often swiftly, on the spot. Other "specialized" items ask teachers to show or represent numbers or operations using pictures or manipulatives, and to provide explanations for common mathematical rules (e.g., why any number can be divided by 4 if the last two digits are divisible by 4).

We believe our measure of teachers' content knowledge bridges the literatures described earlier. It includes the common knowledge often measured within the educational production function literature; however, it also uses lessons from the case study literature on teachers' knowledge to identify and measure the unique skills and capabilities teachers might use in their professional context. By employing this more job-specific measure in the context of an educational production function-type study, we might improve upon prior studies, and examine untested assumptions about the relevance of elementary teachers' mathematical knowledge to student achievement.

Following a review of draft items by mathematicians and mathematics educators both internal and external to the project, we piloted items in California's Mathematics Professional Development Institutes (MPDIs). This pilot allowed us to examine the performance of items – in particular, their level of difficulty vis-à-vis other items, and the strength of their relationship to their underlying construct being measured. Results of this piloting led us to several insights and decisions. Items in the knowledge of students and content category did not meet criteria for inclusion in such a large and costly study.[5]

---

[5] Briefly, many of these items misfit in item response theory models; factor analyses indicated multidimensionality, as some items drew on mathematics knowledge, some on knowledge of students, and some on both jointly; as a set, they were also too "easy" for the average teacher; cognitive tracing interviews suggested teachers' multiple-choice selections did not always match their underlying thinking.

Items in the knowledge of content for teaching category did, however; average reliability for piloted forms ranged in the low .80s with very few misfitting items. Further, specialized factor analyses revealed the presence of a strong general factor in the piloted items (Hill, Schilling & Ball, 2004). Because we had a relatively large pool (roughly 150) of piloted items, we could use information to select SII items with desirable measurement properties, including a strong relationship to the underlying construct, a range of "difficulty"[6] levels, and a mix of content areas.

At the same time we undertook these pilots, we conducted validation work on these items by a) subjecting a subset of items to cognitive tracing interviews and b) comparing items to National Council of Teachers of Mathematics (NCTM) Standards, to ensure that we well covered the domains specified. Results from the cognitive interviews suggest that in the area of content knowledge, teachers produced very few (1.85%) "inconsistent" responses to items, where correct mathematical thinking led an incorrect answer, or incorrect mathematical thinking led to a correct answer (Dean & Goffney, 2004). The content validity check of the entire piloted set indicates adequate coverage across the number, operations, and patterns, functions, and algebra NCTM standards.

The measure of teachers' content knowledge ultimately used in the main study included 30 mathematical knowledge for teaching items on the year 1 through year 3 teacher questionnaires. We balanced items across content domains (13 number items, 13 operations items, 4 pre-algebra items), and specialized (16 items) and common (14 items)

All four problems resulted in our projecting low reliabilities for the number of items that could be carried on the SII TQ. We are continuing to develop theory and measures in an effort to address these results.

[6] "Difficulty" describes the relationship among items, differentiating between those that are easier for the population of teachers as opposed to those that are more difficult. Here, item difficulty is used to ensure that the SII assessment had both easier items – which would allow differentiation among lower-knowledge teachers – and harder items, which would allow the differentiation among higher-performing teachers.

content knowledge. In practice, however, teachers typically answered fewer than 30 items. One reason was that by design, only half the sample answered the first teacher questionnaire.[7] Another reason was that missing data ranges between 5-25% on these items.

We used Item Response Theory (IRT) to handle missing data, linearize scores and provide information about the accuracy of our measures. Teachers' responses were scored in a two-parameter model[8] using Bayesian scoring methods. When a teacher failed to answer more than 25% of CKT-M items on a given questionnaire, we scored that teacher's missing items as "not presented," which does not penalize teachers for skipping items. Otherwise, missing data was scored as incorrect. To confirm the findings presented below, we rescored this data using different methods (i.e., Maximum Likelihood) and handled missing data in different ways (e.g., scored all missing data as not presented). Results were robust to these different methods of computing teacher scores. The reliability of this measure is .88. Finally, the CKT-M measure was calculated for the entire teacher sample (first through fifth grade) as a standardized variable.

In some models below, we also include a content knowledge for teaching English Language Arts measure (CKT-ELA). The objective behind designing the CKT-ELA measure was much the same as in mathematics: to attend to not just the knowledge that adults use in everyday life (i.e., reading text), but also to the specialized knowledge teachers use in classrooms (i.e., determining the number of phonemes in a word;

---

[7] Data collection in year 1 included only 53 of the eventual 115 schools in our sample.

[8] Two-parameter models take into account both the difficulty of an item and the correctness of a response in scoring. Two teachers who both answer 4/5 items correctly, for instance, may have different scores if one correctly answered more difficult items than the other. Missing data in this sample makes 2-parameter models attractive because of this feature. Results in Table 7 were similar with the 1-parameter scoring method.

assessing a piece of text and determining the best question or task to enhance student understanding). The two major content domains included knowledge of word analysis – the process of helping students actually read printed text – and comprehension. The three major teaching domains included knowledge of the content itself, knowledge of students and content, and knowledge of teaching and content. This last category was not represented in the mathematical work, but includes items focused on ways to enhance student learning of particular pieces of text, remediate student problems with text, and so forth. This CKT-ELA measure was constructed through a similar process to the mathematics measure: item-writing by reading educators, experts, and classroom teachers; piloting in California; factor analyses; choosing items for inclusion on the study's teacher questionnaire that balance across the domain map and maximize desired measurement qualities; and IRT scoring. We here use a measure that combines all of the content and knowledge domains and that has a reliability of .92. Details on the construction of this measure can be found in Phelps & Schilling (2004).

School characteristics. The one school characteristic employed in this model is household poverty, or the percentage of households in poverty at the time of the study. This statistic is from the 1990 census.

Models and Statistical Procedures

This paper uses linear mixed models to estimate the influence of student, teacher, and school characteristics on student achievement. As described earlier, student achievement is expressed as gain scores over one year of participation in the study. We used gain scores rather than covariate adjustment models because gain scores are unbiased estimates of student growth (Mullens, Murnane & Willett 1996; Rogosa, Brandt, &

Zimowski 1982; Rogosa & Willett 1985). However, gain score models can be subject to unreliability, and therefore effects of independent variables on gain scores can be underestimated (Rowan, Correnti & Miller 2002). The analyses we report below were produced using PROC MIXED in SAS.

We elected to exclude a number of factors from this model for simplicity of results and discussion. One such factor is instructional practice, as reported on the daily mathematics log. Another is the mathematics curriculum materials used by each school, including whether the school was using the mathematics program recommended by the school reform program. A third is the improvement program selected by the school. Although all are potentially important influences on student achievement, results from initial models suggested their effects are complex – interactive with student background characteristics, for instance, as well as grade level. Participation in a comprehensive school reform program, for instance, had little independent main effect on student achievement; this makes sense, given that the programs participating in this study focus mainly on English Language Arts.

As with any complex study design, there is substantial attrition of students from our sample, and missing data on key instruments. First graders without spring-spring data and third graders without fall-fall assessment data were excluded from the analysis. Teachers were excluded from the analysis when they failed to return any of the three teacher questionnaires, and thus provided us no information on their preparation for teaching, years of experience, or content knowledge for teaching mathematics. When teachers did return questionnaires but did not answer enough content knowledge for teaching (CKT) items to reasonably generate a person-level score, we imputed their score. This resulted in roughly 10% of first grade teachers and 20% of third grade teachers with imputed scores. Teachers who did not log their mathematics instruction had their mean mathematics instructional time and absence rate imputed as well. To preserve cases, we used mean imputation. This is one standard method for dealing with missing cases; one unfortunate side effect is that it does not maintain the actual covariances between variables. However, we do include an indicator (dummy) variable that allows us to examine trends for cases on which there are missing data.

From one perspective, numerous data issues exist in this study, including a small number of students with complete data within each classroom, missing data, a lack of complete alignment between the teacher and student mathematics assessments, and student attrition. With the exception of this last issue, however, all these problems would tend to bias against finding positive teacher/classroom effects in our models, as they make the model less sensitive to actual average classroom growth rate, reduce the amount of observed covariation between inputs and outcomes, and decrease the sensitivity of student assessments to the teacher assessment (see Leinhardt & Seewaldt, 1981, Barr & Dreeben, 1983, and Berliner, 1979 for arguments about overlap). All three suggest that any effects may be stronger than they appear in models.

Results

Table 3 shows pre-standardization sample means and standard deviations for variables included in this analysis. Several variables have substantive interpretations and implications, beginning with student-level descriptives. The average first grader gained nearly 58 points on the mathematics Terra Nova, while the average third grader gained 39 points. This is a two-grade snapshot of the trend toward decelerating growth that appears throughout the study's longitudinal data, and echoes that found in other national studies. Five percent of first graders and four percent of third graders were reported as absent on over 20% of the logs for which they were the designated target student. Finally, roughly 70% of our student sample is non-Asian students of color.

Several teacher-level descriptives also stand out. Because we averaged reports of mathematics methods and content courses, and because teachers report such courses as ranges (e.g., 1-3 courses, 4-6 courses), the measure representing these reports has no easy substantive interpretation. However, it may help the reader to know that 12% of teachers reported never having taken a mathematics content or methods course, 15% of teachers reported taking between 1 and 3 such courses, and 27% of teachers reported taking between 2 and 6 courses. In many colleges of education, the mathematics methods course is taught by education school faculty, and typically covers the use of manipulatives and other representations for content, problem solving, classroom organization, and designing and teaching math lessons. The mathematics content courses are often taught by a member of the mathematics department and usually cover mathematical topics in the K-6 curriculum — whole numbers and fractions, place value, probability, geometry, combinatorics, and, often, problem solving. Some other required mathematics content courses may be the same as those taken by mathematics majors.

Nearly 90% of our sample is certified, and the average teacher is in her twelfth year of teaching. The average teacher reports spending nearly an hour per day on mathematics instruction: 55.6 minutes for first graders and 50.3 minutes for third graders. These figures comprise all days on which mathematics could have been recorded as taught, including days on which mathematics was not taught due to an assembly, field trip, test preparation, or similar interruption. Finally, the average teacher reported herself absent on 5-6% of logging days, or for roughly 9 days of a 180 day school year. This figure doubtlessly includes professional development days in addition to actual absences.

The sole school level variable included in this analysis, household poverty, shows that at the average school roughly one in five students lives in a household below the poverty line. This measure is similar to the student-level SES measure, and captures the additional effect of poverty concentration within schools on student achievement.

Tables 4 and 5 show the correlations among the teacher preparation, experience, and knowledge variables. The size and strength of relationships is relatively constant at these two grades, and several relationships stand out. Years of experience in teaching is modestly correlated with certification as well as methods and content courses. This is consistent with the observation that some proportion of our sample is newly emergency-credentialed teachers with incomplete formal teacher training or certification, and that teachers continue to take mathematics courses as part of graduate-level study. Teachers' mathematical content knowledge for teaching, on the other hand, is not significantly correlated with any of the teacher preparation or experience variables at grade 1, and only mildly with teacher certification at grade 3. We cannot make any conclusions about causation from these correlations, but this finding suggests that neither ensuring teacher certification nor increasing teachers' subject-matter or methods coursework, two common approaches to improving teacher quality, ensures a supply of teachers with strong content knowledge for teaching mathematics. Finally, teachers' mathematics CKT and language arts (ELA) CKT measures are correlated, but not as strongly as one might expect: .39 and .37 in the first and third grades, respectively.

Table 6 shows results of an unconditional model decomposing the variance in student gain scores into that which lies among schools, teachers, and students in the third grade data. The largest amount of variance, 85%, lies among students. This statistic is in line with that found in other studies and makes sense: students' gain scores are influenced by native intelligence, motivation, behavior, personal educational history, and family support for educational outcomes. By comparison, only a small amount of the remaining variance lies between teachers – roughly 8% for first grade, and 2% for third grade. This estimate is probably artificially low because of unreliability in measurement of student gains and the small number of students per classroom, yet it also reflects the common-sense assumption that teachers are but one influence on student gains over the course of a year. To determine whether teacher-level effects can be further modeled, we conducted likelihood ratio test for variance components; this test rejected the null hypothesis that there is no meaningful variance between teachers. Six percent and 7% of the variance lies among schools in the first and third grades, respectively.

Table 7 shows two models each for first and third grade. All independent variables were standardized before entry into these models, making coefficients easily interpretable as the effect of a one standard deviation increase in each independent variable on the dependent variable, student gains over one year. Student level variables, which remain the same in both models, are the strongest predictor of gain scores. Initial mathematics Terra Nova scores are strongly and negatively related to gain scores. Students who perform well at the initial assessment, in other words, tend to regress to more average performance on the second assessment. Family socio-economic status (SES) is also a strong predictor of gain scores; for every one standard deviation increase in socio-economic status, students gain an additional 2 to 4 points. Missing family SES data is not related to student gains at the first grade, but is negatively related to student gains at the third grade, where the proportion of missing data for our sample was higher. This suggests that families of third graders who did not respond to the phone interview had students who gained less over the course of the year. Female students saw a two-point gain relative to males in the third grade but not the first. Non-Asian minority students have lower gain scores in the first grade and, more marginally ($p=.11$), in the third grade. Students reported absent on more than 1/5 of days they were a subject of logging (high absence) also gained less than students with lower absence rates in the third grade model; this effect was close to significant ($p<.10$) in the first grade model as well. Though these models are not fully enough specified to explore the subtle effects of race, culture, and SES on student achievement, the results are consistent with other research in this arena, and we are satisfied that key student covariates have been captured, thus allowing the teacher-level modeling we discuss below.

Teachers' content knowledge for teaching mathematics is a significant predictor of student gains in both models at both grade levels. The effect is strongest in Model 1, where students accrue an additional two and a quarter points on the Terra Nova for every standard deviation difference in teachers' mathematics content knowledge. Expressed as a fraction of average monthly student growth in mathematics, this translates to roughly 1/2 to 2/3 of a month of additional growth per standard deviation difference on CKT-M. CKT-M is the strongest teacher-level predictor in these models, larger than teacher background variables, and greater than the average time spent on mathematics instruction each day. In third grade, its effect size rivals that of SES and students' ethnic and gender affiliation and in the first grade models, the size is not far off. This suggests that knowledgeable teachers can positively and substantially affect student learning of mathematics, and the size of this effect is, at least in this sample, in league with the effects of student background characteristics.

An important question is whether the effect of teachers' content knowledge on growth in student achievement is linear – that is, whether the gain of slightly over two points per standard deviation of teacher CKT is constant across the range of teacher knowledge. Perhaps only the most knowledgeable teachers deliver highly effective mathematics instruction; alternatively, it may be that only the least knowledgeable teachers have any effect on students' mathematics achievement. To investigate this question, we divided teachers into deciles by their CKT-M score, with the lowest decile (1) representing the least knowledgeable teachers. We replaced the linear measure of CKT-M in Model 1 with this new ten-category demarcation of teachers, and show the results – estimated student gains per CKT-M decile – in Figures 2 and 3. Teachers in the lowest two deciles (0-20%) of the first grade CKT distribution taught students who gained, on average, nearly 10 fewer points than students in the highest category, which was the referent. However, above the lowest two deciles there appears little systematic relationship between increases in teacher knowledge and student gains. A statistical test for difference of means (in SAS, lsmeans test) across categories confirms that significant differences occur only between the lowest 20% of teachers and other categories. In the third grade data (Figure 3), the significance test suggests teachers in the first three deciles (0-30%) significantly impact their students' achievement vis-à-vis the top four deciles.

Despite success in identifying a positive relationship between mathematical knowledge for teaching and student gain scores, the possibility remains that general knowledge of or aptitude for teaching, not content-specific knowledge for teaching, has produced this finding. We have no measure of general knowledge of teaching, and cannot directly answer this claim. However, we do have a measure of content knowledge for teaching English Language Arts that is similar in intent to the CKT-M measure, but intended to measure teachers' knowledge of and ability to teach word analysis and comprehension. If the ELA and mathematics measures both draw heavily on general knowledge of teaching, they should be moderately to highly correlated, and should share the positive relationship to student achievement seen in Model 1. In Model 2, we include these CKT-ELA measures and find that although the CKT-ELA measure is positively related to student gains at the first and third grade, it is not statistically significant. Further, it has only a small effect on the absolute size and significance of the CKT-M variable. This suggests that the effect of teachers' knowledge on student achievement is at least content-specific, and in mathematics, reflects more than just having more generally knowledgeable teachers in classrooms.

Our models hold other significant or near-significant findings. The average length of a teachers' mathematics lesson was significantly related to third grade student gains, with a one-standard deviation in daily mathematics lesson length – about 14 minutes – yielding an additional 1.8 points. This translates to roughly an additional 2 weeks of instruction per year for a classroom that receives the additional 14 minutes per day. Mathematics preparation, or the average number of content and methods courses taken in pre-service or graduate training, positively predicts student gains in the third grade, but lies just outside of traditional significance (p = .06). We are unsure how to interpret this finding, both because it is not consistent across grade levels, and because the lack of a correlation between content knowledge and math preparation courses suggests these measure two separate constructs. The effects of another commonly argued policy solution, teacher certification, were non-existent in this particular sample of teachers and students. Although certification is mildly related to teachers' knowledge of content in the third grade (Table 5), it exerts no independent influence over student gain scores. This may reflect a true null effect, or a reflection of Table 5's suggestion that non-certified teachers have taken as many math methods and content courses as certified teachers. Thus non-certified teachers may be en route to traditional certification, or transfers into new schools from other states (Darling-Hammond, Berry & Thoreson, 2001) or mathematics-intensive professions. It could also reflect the fact that certification requirements vary across the states included in the study.

Years of teaching experience, measured linearly, shows no relationship to first grade student achievement, and a marginally significant (p = .11) positive relationship in the third grade. Some studies, however, have suggested that it is teachers in the first several years of their career who negatively impact student achievement. We created a dummy variable representing teachers in their first or second years of teaching and entered it into the models in place of the linear measure. The significance of this variable in the third grade model did not change, but this measure of novice teachers did become marginally significant in the first grade model (b= -5.3, p <.10).

We checked these models in several ways: adding and deleting variables to check for model stability; using pre-on-post models rather than gain score models; creating dummy variables to check for linearity. Overall significance of key variables held firm, and residuals were normally distributed.

Conclusion

This analysis has clear limitations, including the sample of students, missing data, and a lack of alignment between our measure of teachers' mathematical knowledge and student achievement. Because many of these problems would bias the coefficients on our content knowledge variable toward zero, we feel confident that the positive effect we see on this variable is robust and, if anything, underestimated. However, we are less confident in any borderline or null results, such as on the teacher preparation measure, and thus focus our discussion mainly on the results from the content knowledge variable.

We found teachers' mathematical knowledge for teaching positively predicts student gains in the first and third grade. We were modestly surprised to see the first grade effect, since we had expected the CKT-M measure to have its effects mainly at grades with more complex content – e.g., at grade levels where multidigit addition or multiplication, functions, fractions, and decimals was taught. That it also positively affects student gains in the first grade suggests teachers' content knowledge plays a role even in the teaching of very elementary content. Many kindergarten and first grade teachers explain their choice of grade level by referencing both their love of young children and lack of mathematics knowledge. However, our analysis suggests that mathematical knowledge for teaching is important, even at this grade level, in our sample of schools.

Important in understanding these results is that ours is a measure of mathematical knowledge for teaching, and not merely of teachers' computational facility or course-taking. Although scholars from John Dewey (1904) to Joseph Schwab (1964) to Lee Shulman have observed that teachers' responsibilities for subject matter require special sensitivities to the content, the nature of this special knowledge has not been elaborated. Consequently, it has been difficult to measure reliably or validly on a large scale. Our work built on these scholars' theories about relationships of subject matter and pedagogy to design a measure of teachers' mathematical knowledge for teaching, and we can report here that this more task-sensitive measure is positively related to student achievement.

This modifies findings from earlier studies exploring the effect of teachers on student achievement. It confirms one critique of the process-product literature, namely, that the subject does matter, and helps envision a new generation of process-product studies designed to answer questions about how teachers' mathematical behavior – in particular their classroom explanations, representations, and interactions with students' mathematical thinking – might affect student outcomes. It also informs findings from the educational production function literature, first by pointing out that a direct measure of teachers' content knowledge for teaching trumps proxy measures such as courses taken or experience, and then by suggesting that measures of teacher knowledge be at least content-specific, and even better, specific to the knowledge used in teaching children.

Our findings both support and challenge recent policy initiatives. If successful, efforts to improve teachers' mathematical knowledge through content-focused professional development and pre-service programs targeted toward low-performing/high poverty schools can raise student achievement, as intended. Such programs include California's Mathematics Professional Development Institutes, the National Science Foundation/Department of Education's Math-Science Partnerships, and many other local efforts throughout the U.S. Yet our results suggest that those who may benefit most are teachers in the lowest third of the distribution of knowledge, and that efforts to recruit teachers into professional development and pre-service coursework might focus most heavily on those with weak subject matter knowledge. However, without ways to differentiate and select such teachers, and without strong incentives for bringing such teachers into content-focused professional development, the intended effects of these major programs may be lost. Moreover, without conceptual and analytic tools for examining whether and what teachers learn from such professional development, efforts to develop the quality and effectiveness of programs designed to improve teaching will be impeded.

Another key question for these data involves equity, namely the intellectual resources available to students across race and socio-economic status (see Cohen, Raudenbush & Ball, 2003 for a discussion of such resources). In the first grade, teachers' mathematical knowledge for teaching is distributed fairly evenly across students of different socio-economic status, but there is a negative relationship between student minority status ($r = -.16$, $p < .01$) and teachers' mathematical knowledge for teaching. In the third grade, the relationship between student SES and teacher knowledge becomes significant ($r = .11$, $p < .05$) and the relationship between minority status and teacher knowledge increases ($r = -.26$, $p < .0001$). These results are similar to those found elsewhere with other samples of schools and teachers (Hil & Lubienksi, under review; Loeb & Reininger, 2004). This problem is particularly pressing if the relationship of teachers' mathematical knowledge to instructional quality is nonlinear, as this analysis suggests. A portion of the achievement gap on the National Assessment of Educational

Progress and other standardized assessments may result from teachers with less mathematical knowledge teaching more disadvantaged students. Closing this gap, then, may require improving the quality of mathematics content knowledge among teachers across settings in the U.S.

Three additional lines of inquiry grow naturally from this analysis. The first is examining the effect of mathematics instructional methods and student curriculum materials (texts) on student performance. A key component of this analysis will be examining interactions between teacher knowledge and instructional method/uses of texts. A second line of inquiry should parse more precisely different theoretically- and empirically-grounded distinctions in content knowledge for teaching, and to investigate their relationships, separately and in combination, to student achievement. The analyses reported here do not make such distinctions, and it is possible that effects may differ across types of knowledge — common (CCK), specialized knowledge of content (SKC), as well as knowledge of students and content (KSC) and knowledge of content and teaching (KCT) (see Hill, Schilling & Ball, 2004).

Finally, a third line of inquiry should focus on investigating whether and how the instructional practice of mathematically knowledgeable and less knowledgeable teachers differs. Teachers do not improve student learning simply by scoring well on our multiple-choice assessment. However, what knowledgeable teachers do in classrooms – or how knowing mathematics affects instruction – has yet to be studied and analyzed. Does teachers' knowledge of mathematics affect the decisions they make? Their planning? How they work with students, or use their textbooks? How they manage student confusions or insights, or how they explain concepts? Previous research on teachers' pedagogical content knowledge suggests knowledgeable teachers may provide better mathematical explanations, construct better representations, better "hear" student methods and have a clearer understanding of the structures underlying elementary mathematics and how they connect (e.g., Ball 1993; Borko et al 1992; Carpenter et al 1989; Leinhardt & Smith 1985; Ma 1999; Thompson & Thompson 1994). However, analyzing the practice of knowledgeable teachers may also surface new aspects of the mathematical knowledge that matters for teaching: how mathematical and everyday language are bridged, for example, or how representations are deployed, or how numerical examples are selected. Ongoing research on teaching, on students' learning and performance, and on the mathematical demands of high quality instruction can contribute to increasing precision in our knowledge of the role of content knowledge in teaching.

References

Ball, D. L. (1988). Knowledge and reasoning in mathematical pedagogy: Examining what prospective teachers bring to teacher education. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Ball, D.L. (1990). The mathematical understandings that prospective teachers bring to teacher education. Elementary School Journal, 90, 449-466.

Ball, D. L. (1991). Teaching mathematics for understanding: What do teachers need to know about subject matter? In M. Kennedy (Ed.), Teaching academic subjects to diverse learners (pp. 63-83). New York: Teachers College Press.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. Journal of Teacher Education, 51, 241-247.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), Multiple perspectives on the teaching and learning of mathematics (pp. 83-104). Westport, CT: Ablex.

Ball, D. L., & Bass, H. (2003). Making mathematics reasonable in school. In G. Martin (Ed.), Research compendium for the Principles and Standards for School

Mathematics. (pp. 27-44) Reston, VA: National Council of Teachers of Mathematics.

Ball, D., Camburn, E., Correnti, R., Phelps, G., & Wallace, R. (1999). New tools for research on instruction: A web-based teacher log. Working paper, Center for Teaching Policy. Seattle: University of Washington.

Ball, D.L. Corey, D., & Harrison, D. (2004) Investigating school time devoted to mathematics and English Language Arts instruction. Paper presented at American Educational Research Association Annual Meeting, San Diego CA.

Barr, R. & Dreeben, R. (1983) How schools work. Chicago: University of Chicago Press.

Begle, E. G. (1972). Teacher knowledge and student achievement in algebra (SMSG Rep. No. 9). Palo Alto, CA: Stanford University.

Begle, E. G. (1979). Critical variables in mathematics education: Findings from a survey of the empirical literature. Washington, DC: Mathematical Association of America and National Council of Teachers of Mathematics.

Benson, G. (2002) Study of Instructional Improvement School Sample Design. University of Michigan, Ann Arbor: Institute for Social Research.

Berliner, D. (1979) Tempus Educare. In Research on Teaching: Concepts, Findings, and Implications, P. Peterson and H. Walberg (Eds.), pp. 120-135. Berkeley, Calif.: McCutchan.

Boardman, A.E., Davis, O.A.,& Sanday, P.R. (1977) A simultaneous equations model of the educational process. Journal of Public Economics 7, 23-49.

Borko H., Eisenhart, M. Brown, C. A., Underhill, R. G., Jones, D. & Agard, P.C. (1992) Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? Journal for Research in Mathematics Education 23, 194-222.

Bruner, J. (1960). The process of education. Cambridge, MA: Harvard University Press.

Carpenter, T. P., Fennema, E., Peterson, P.L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. American Educational Research Journal 26, 499-531.

Cohen, Raudenbush & Ball (2003) Resources, Instruction and Research. Educational Evaluation & Policy Analysis 25, 119-42.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001) Does teacher certification matter? Evaluating the evidence. Educational Evaluation and Policy Analysis (23), 57-77.

Dean, C. A. & Goffney, I. M. Technical report on the validity of CKT-M items. [Name of Institution]

Dewey, J. (1902). The child and the curriculum. Chicago: University of Chicago Press.

Ferguson, R.F. (1991) Paying for public education: New evidence on how and why money matters. Harvard Journal on Legislation, 28, 458-498.

Greenwald, R., Hedges, L.V., & Laine, R.D. (1996) The effect of school resources on student achievement. Review of Educational Research, 6, 361-396.

Grossman, P. L. (1990). The making of a teacher : Teacher knowledge and teacher education. New York: Teachers College Press.

Hanushek, E.A. (1972) Education and race: An analysis of the educational production process. Lexington, MA: D.C. Heath & Co

Hanushek, E.A. (1981) Throwing money at schools. Journal of Policy Analysis and Management 1, 19-41.

Hanushek, E.A. (1996) A more complete picture of school resource policies. Review of Educational Research 66, 397-409.

Harbison, R.W. & Hanushek, E.A. (1992) Educational performance for the poor: Lessons from rural northeast Brazil. Oxford, England: Oxford University Press.

Hill, H.C., Schilling, S.G., & Ball, D.L. (in press) Developing Measures of Teachers' Mathematics Knowledge for Teaching. Elementary School Journal.

Hill, H.C. & Lubienski, S.T. (in progress) Teachers' Mathematics Knowledge for Teaching and School Context: A Study of California Teachers. Ann Arbor, MI: University of Michigan.

Leinhardt , G., & Seewaldt, A.M. 1981. Overlap: What's Tested, What's Taught. Journal of Educational Measurement 18 (2): 85-95.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. Journal of Educational Psychology, 77, 247-271.

Loeb, S. & Reininger, M. (2004) Public policy and teacher labor markets: What we know and why it matters. East Lansing, MI: The Education Policy Center at Michigan State University.

Ma, L. (1999). Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States. Mahwah, NJ: Erlbaum.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. Economics of Education Review, 13, 125-145.

Mullens, J. E., Murnane, R. J., Willett, J. B. (1996) The contribution of training and subject matter knowledge to teaching effectiveness: a multilevel analysis of longitudinal evidence from Belize. Comparative Education Review 40, 139-57.

Rogosa, D., Brandt, D., & Zimowski, M. (1982) A growth curve approach to the measurement of change. Psychological Bulletin, 92, 726-748.

Rogosa, D.R. & Willett, J.B. (1985) Understanding correlates of change by modeling individual differences in growth. Psychometrika, 50, 203-228.

Rowan, B., Chiang, F., & Miller, R.J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. Sociology of Education, 70, 256-284.

Rowan, B., Correnti, R. & Miller, R. J. (2002) What large-scale survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of Elementary Schools. Teachers College Record, 104, 1525-1567.

Rowan, B., Harrison, D., & Hayes, A. (in press) Using instructional logs to study mathematics curriculum and teaching in the early grades. Elementary School Journal.

Schwab, J. (1961/1978). Science, curriculum, and liberal education: Selected essays. Chicago: University of Chicago Press.

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

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57, 1-22.

Strauss, R.P. & Sawyer, E.A. (1986) Some new evidence on teacher and student competencies. Economics of Education Review, 5, 41-48.

Summers, A.A. & Wolfe, B.L. (1977) Do schools make a difference? American Economic Review, 67, 639-652.

Thompson, P. & Thompson, A. (1994) Talking about rates conceptually, Part I: A teachers' struggle. Journal for Research in Mathematics Education 25, 279-303.

Wayne, A.J. & Youngs, P. (2003) Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73, 89-122.

Wilson, S. M. & Wineburg, S. (1988). Peering at history through different lenses: The role of disciplinary perspectives in teaching history. Teachers College Record, 89, 525-39.

Wilson, S. M., Shulman, L. S., & Richert, A. (1987). 150 different ways of knowing: Representations of knowledge in teaching. In J. Calderhead (Ed.), Exploring teachers' thinking (pp. 104-124). Sussex, England: Holt, Rinehart & Winston.

Table 1

SII vs. ECLS students

	SII (n=1,616)	**ECLS** (n=21,116)
Household income		
UNDER $5,000	3.5%	3.3%
$5,000 - $9,999	8.3%	4.2%
$10,000 - $14,999	9.8%	7.7%
$15,000 - $19,999	9.5%	6.8%
$20,000 - $24,999	9.0%	7.9%
$25,000 - $29,999	8.4%	6.4%
$30,000 - $34,999	7.6%	7.0%
$35,000 - $39,999	6.6%	5.6%
$40,000 - $49,999	9.1%	10.3%
$50,000 - $74,999	18.9%	20.0%
$75,000 - $99,999	5.6%	9.5%
$100,000 - $199,999	4.3%	8.8%
$200,000 or more	3.3%	1.95%
Mother's educational background	n=1,840	n=19,809
Did not complete high school	18.3%	14.3%
High school diploma or equivalent	34.7%	30.6%

Some college or vocational school	34.9%	31.7%
Bachelor's degree	9.3%	14.6%
Master's degree or attended professional school	2.4%	5.9%
Ph.D. or other advanced degree	0.4%	1.4%
Father's educational background	n=1,205	n=16,066
Did not complete high school	14.4%	11.2%
High school diploma or equivalent	34.1%	26.0%
Some college or vocational school	29.1%	20.8%
Bachelor's degree	12.1%	13.0%
Master's degree or attended professional school	4.3%	5.7%
Ph.D. or other advanced degree	1.0%	3.2%
Family structure	n=1,900	n=18,962
Biological mother/father present in household	52.2%	63.8%
Parent with stepparent or partner	6.5%	9.5%
Single parent	40.7%	22.6%
Student race	n=2,130	n=21,190
White	26.2%	57.0%
Black	47.5%	16.4%
Hispanic	14.3%	18.9%
American Indian/Alaskan Native	0.5%	1.8%

Asian or Pacific Islander	5.1%	3.4%
Hispanic	16.4%	20.6%
Other	4.3%	2.4%

Table 2

Instrument response rates

	2000-2001		2001-2002		2002-2003	
	sample / completed rate	Pct.	sample / completed rate	Pct.	sample / completed rate	Pct.
Self-Administered Questionnaires						
Teacher Questionnaire (TQ)	2874/1806	69%	4043/2969	73%	3751/2861	76%
Teacher Logs						
Teacher sample-math log	178/172	97%	570/519	91%	n/a[b]	
Completed logs –filtered [a]	9025/8216	91%	31414/28560	91%	n/a	
Parent Interview						
Parent Questionnaire (PQ)	2343/1999	85%	3777/2877	76%	n/a	
Student Instruments						
Terra Nova (TN) - Fall	1289/1247	97%	3845/3690	96%	4868/4638	95%
Terra Nova (TN) - Spring	2313/2220	96%	5080/4897	96%	4743/4595	97%

[a] Log samples filtered by teacher refusal, student move-out, student ineligible, and parental refusal.

[b] n/a indicates data from this year not used in this paper

Table 3

Sample means and standard deviations

		Grade 1			Grade 3		
Label	Description	Mean	SD	N	Mean	SD	N
Student variables							
Average gain	Spring K – Spring 1st Fall 3rd– Fall 4th	57.6	34.6	1190	39.4	33.1	1773
Initial math score	Initial math Terra Nova score	466.6	41.5	1190	563.7	36.2	1773
SES	Family socio-economic status	-.01	.74	1190	-.05	.66	1773
SES missing	No data on family socio-economic status	.07	.26	1190	.23	.42	1773
High absence	Marked 1 if student's absence rate exceed 20%	.05	.22	1190	.04	.09	1773
Female	Marked 1 if student is female	.51	.50	1190	.53	.50	1773
Minority	Marked 1 if student is non-Asian minority	.68	.47	1190	.70	.46	1773
Teacher/classroom variables							

Math methods & content	Math methods and content courses taken	2.56	.95	334	2.50	.91	365
Certified	Marked 1 if teacher is certified	.89	.31	334	.90	.25	365
Years experience	Years experience reported in Year 2 of study	12.21	9.53	334	12.85	9.45	365
CKT-Mathematics	Content knowledge for teaching mathematics	.03	.97	334	.05	.89	365
CKT-M Missing	Missing content knowledge for teaching mathematics	.09	.29	334	.19	.39	365
CKT-ELA	Content knowledge for teaching English Language Arts	.14	.74	334	.07	.64	365
CKT-ELA Missing	Missing content knowledge for teaching English Language Arts	.08	.27	334	.18	.38	365
Math lesson length	Average length in minutes of mathematics class	55.6	13.4	334	50.3	14.4	365
Teacher absence rate	Percent of logs on which teacher reports own absence	.05	.22	334	.06	.05	365
Log data missing	No information on math lesson length, teacher	.06	.24	334	.10	.31	365

	absence or student absence						
Pct class minority	Percent minority in a classroom, initial time point	.47	.32	334	.64	.35	365
<u>School-level variables</u>							
Household poverty	Percent of households in poverty	.18	.13	115	.19	.14	115

Table 4

Correlations between teacher preparation, experience, and mathematical knowledge for teaching – 1st grade

	Math methods & content	Certified	Years experience	CKT-Math	CKT-ELA
Math methods & content	1.0	.10	.18*	.00	-.07
Certified		1.0	.20**	.07	.04
Years experience			1.0	.00	.01
CKT-Mathematics				1.0	.39**
CKT-ELA					1.0

* Significant at $p < .05$

** Significant at $p < .001$

Table 5

Correlations between teacher preparation, experience, and mathematical knowledge for teaching – 3st grade

	Math methods & content	Certified	Years experience	CKT-Math	CKT-ELA
Math methods & content	1.0	.03	.19**	-.08	-.05
Certified		1.0	.15*	.11*	.02
Years experience			1.0	-.09	.05
CKT-Mathematics				1.0	.37**
CKT-ELA					1.0

* Significant at $p < .05$

** Significant at $p < .001$

Table 6

Variance Components

	Grade 1	Grade 3
Teachers	99.2	24.4
Schools	77.4	79.3
Residual	1028.3	990.27
Total	1204.9	1093.97
AIC	11774.8	17386.3

Table 7

Student gain score models

	Model 1		Model 2	
	Grade 1	Grade 3	Grade 1	Grade 3
Intercept	57.6 (1.31)	39.3 (.97)	57.6 (1.31)	39.3 (.97)
Student variables				
Initial math score	-19.5*** (1.21)	-17.0*** (.84)	-19.5*** (1.21)	-17.1*** (.84)
SES	3.96*** (.94)	2.13** (.76)	3.95*** (.94)	2.12** (.76)
SES missing	.15 (.72)	-1.80* (.73)	.15 (.73)	-1.80* (.73)
Female	-.55 (.87)	1.80** (.70)	-.56 (.87)	1.79** (.69)
Minority	-4.15** (1.43)	-1.86 (1.15)	-4.14*** (1.43)	-1.84 (1.15)
High absence	-1.51 (.88)	-.74* (.38)	-1.55 (.88)	-.74* (.38)
Teacher/classroom variables				
Math methods & content	.53 (1.00)	1.64 (.92)	.55 (1.01)	1.70 (.92)
Certified	.23 (.89)	-.34 (.73)	.24 (.90)	-.33 (.72)
Years experience	.72 (1.14)	1.02 (.64)	.71 (1.14)	.95 (.66)
Background variables missing	-.22 (.96)	-.61 (.81)	-.20 (.95)	-.57 (.80)
CKT-Mathematics	2.22* (.91)	2.28** (.75)	2.12* (1.00)	1.96** (.77)
CKT-ELA			.26	.82

			(1.18)	(.87)
CKT missing	-.64	-.31	-.64	-.22
	(1.25)	(1.00)	(1.27)	(1.02)
Math lesson length	-.11	1.77*	-.11	1.82*
	(1.04)	(.87)	(1.05)	(.88)
Teacher absence rate	-1.01	-.37	-1.00	-.36
	(.92)	(.88)	(.94)	(.88)
Log data missing	-1.80*		-1.81*	
	(.91)		(.91)	
Pct class minority	2.29	-2.22	2.34	-2.20
	(1.37)	(1.28)	(1.41)	(1.28)
School-level variables				
Household poverty	-1.60	-1.59	-1.60	-1.64
	(1.33)	(1.02)	(1.33)	(1.02)
Variance components				
Teacher	80.63	13.8	84.6	14.7
School	82.40	53.2	79.88	52.6
Residual	730.89	774.5	730.65	774.11
AIC	11342.4	16836.0	11340.1	16833.6

* Significant at p < .05

** Significant at p < .01

*** Significant at p < .001

Figure 1

Examples of Items Measuring Content Knowledge for Teaching Mathematics

1. Mr. Allen found himself a bit confused one morning as he prepared to teach. Realizing that ten to the second power equals one hundred ($10^2 = 100$), he puzzled about what power of 10 equals 1. He asked Ms. Berry, next door. What should she tell him? (Mark (X) ONE answer.)

a) 0

b) 1

c) Ten cannot be raised to any power such that ten to that power equals 1.

d) -1

e) I'm not sure.

2. Imagine that you are working with your class on multiplying large numbers. Among your students’ papers, you notice that some have displayed their work in the following ways:

Student A	Student B	Student C
35 x 25 125 +75 875	35 x25 175 +700 875	35 x 25 25 150 100 +600 875

Which of these students would you judge to be using a method that could be used to multiply any two whole numbers?

	Method would work for all whole numbers	Method would NOT work for all whole numbers	I’m not sure
a) Method A	1	2	3
b) Method B	1	2	3
c) Method C	1	2	3

Figure 2

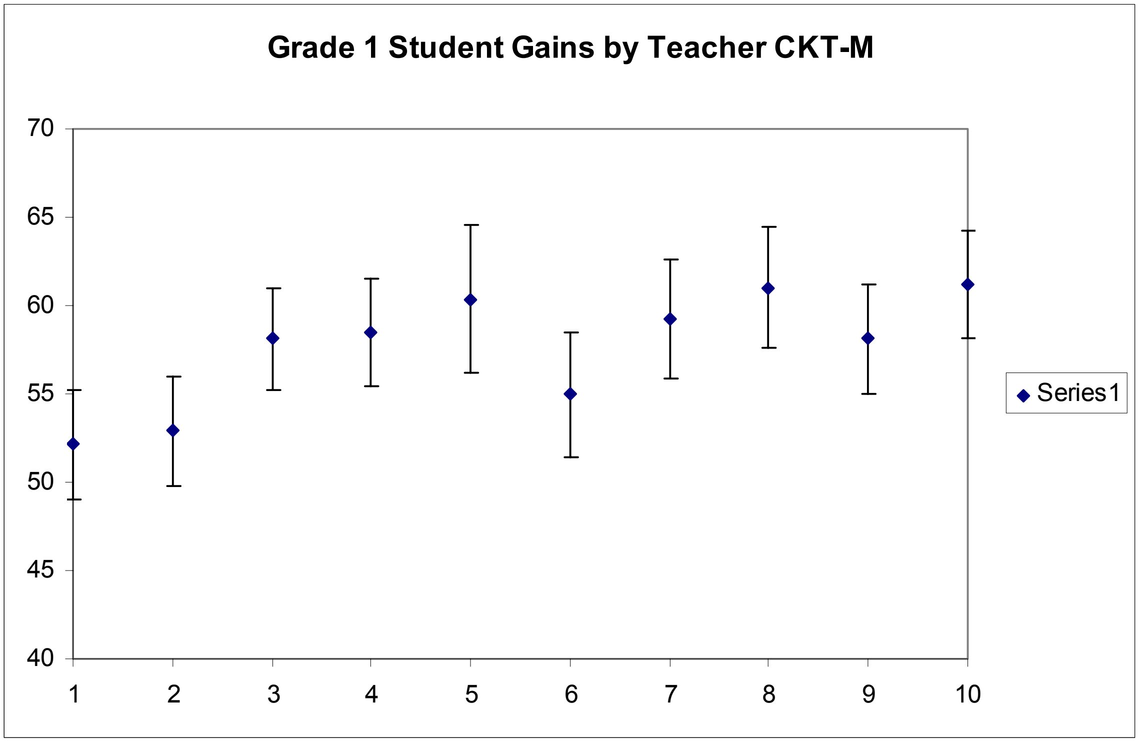
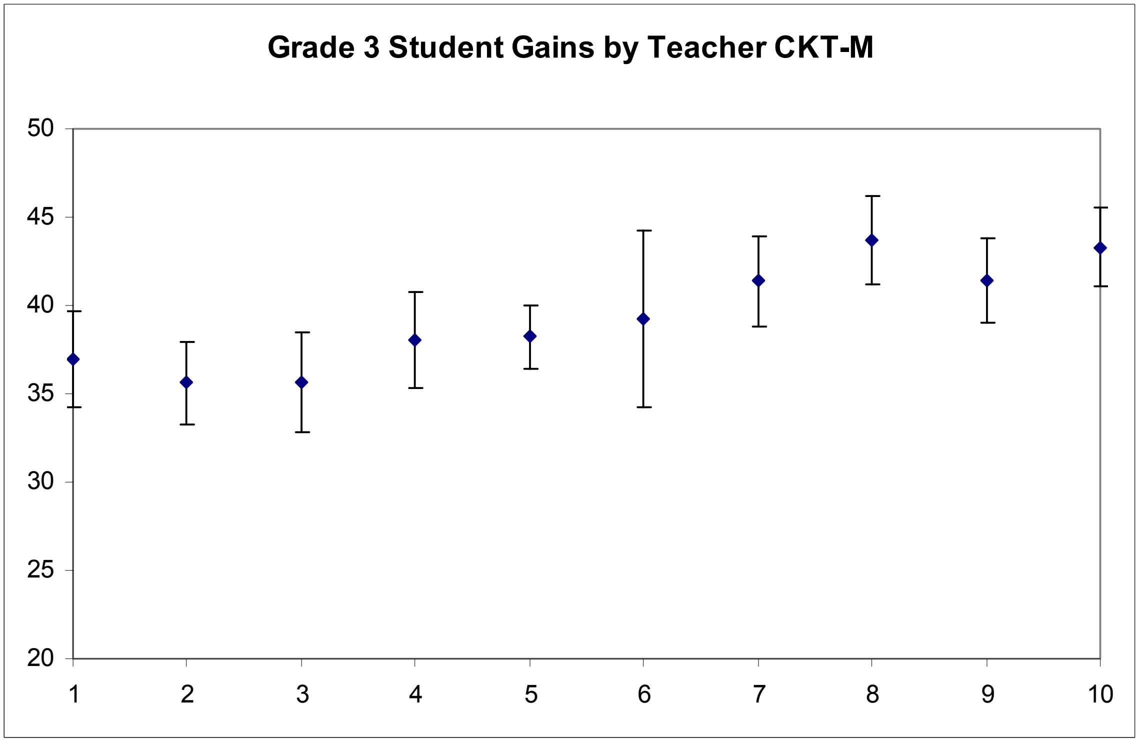

Grade 1 Student Gains by Teacher CKT-M
70
65
60
55
50
45
40
1
2
3
4
5
6
7
8
9
10
◆ Series1

Figure 3


Grade 3 Student Gains by Teacher CKT-M
50
45
40
35
30
25
20
1
2
3
4
5
6
7
8
9
10