A Comparison of Two Prominent Instructional Approaches to the Teaching and Learning of Multi-digit Computation

by

Delena Marie Harrison

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Educational Studies)
in the University of Michigan
2013

Doctoral Committee:

Professor Deborah Loewenberg Ball, Chair
Professor Hyman Bass
Professor Brian Rowan
Professor Stephen Raudenbush, University of Chicago

Dedication

To Jeff, Benjamin, Anna, and Claudia — my family — who have been so patient when I've said, "I need to work on my dissertation," but still have no idea what I've been doing all this time

To my parents, Sandra and Lenn, for their support from beginning to end

To Abby and in memory of Tiffany, who have been the best little pets and perfect company as I worked on this dissertation

Acknowledgements

I have enjoyed working on this dissertation and working with the data from the Study of Instructional Improvement. This work has afforded many more experiences than are evident in this dissertation.

First, I would like to thank the principal investigators for The Study of Instructional Improvement — Deborah Loewenberg Ball, David K. Cohen, and Brian Rowan — for providing financial support and graduate research support early in this work. This project gave me the chance to work with many other researchers — Sally Atkins-Burnett, Eric Camburn, Heather C. Hill, Doug Corey, Geoffrey Phelps, Charles Vanover, Stephen Schilling, and especially my officemates, Keisha Ferguson, Andy Hayes and James Taylor. My work with this project came at a time when I was taking methods courses and this project gave me opportunities to try out these "new" methods. I would also like to thank the School of Education for several years of financial support.

Second, I would like to thank my committee, Deborah Ball, Brian Rowan, Stephen Raudenbush, and Hyman Bass, for their comments along the way, for their patience as I took breaks in my work to have three children, and for their constant belief that I could finish this dissertation.

Third, I would like to thank several groups of people who contributed knowledge along the way. Ravin Pan, Mark Hoover-Thames,

and others helped with a validation study on my use with the achievement items. Stephen Schilling shared his expertise in psychometric analysis and consulted on my use of the achievement items. Stephen Raudenbush shared his expertise over and over as I tried to fairly use the SII data for a secondly investigation. Graeme Tank shared his Ruby programming skills when I was less than thrilled to run the many t-tested needed for testing balance on the propensity scores. Even with it programmed to run nearly automatically, it ran for over a week! So thank you Graeme! I would like to thank Dr. Furong Gao, from CTB McGraw-Hill, with her help in trying share TerraNova item parameter information and a draft of a nondisclosure agreement. Thanks also goes to University of Michigan's legal department for help with this agreement. I also thank Brady West and Carrie Hsoman with CSCAR for consulting on the propensity score analysis.

Fourth, I would like to thank my special friends, Guanglei Hong, Ji-Soo Kim, and Naomi Norman for talking, listening, and thinking with me at various stages of my work.

Fifth, I would like to thank my chair, Deborah Ball, for her persistence, patience, timely comments, coaching, sensitivity and understanding. I have needed each of these qualities from her at different times. Thank you! Without you, it would not be done.

Finally, a special thanks goes to my family. Thank you for sticking with me till the end.

Table of Contents

List of Tables

List of Figures

List of Appendices

ABSTRACT

A Comparison of Two Prominent Instructional Approaches

to the Teaching and Learning of Multi-digit Computation

by

Delena Marie Harrison

Chair: Deborah Loewenberg Ball

This dissertation compares two approaches commonly used to teach multi-digit computation in second through fifth grades. The two instructional approaches under investigation differ in their emphasis on aspects of mathematical proficiency, often referred to as *procedural* and *conceptual* knowledge. The question—of which approach should be preferred— has been the center of debates in mathematics instruction for over 20 years. Instruction that differs in its emphasis on procedural and conceptual knowledge is thought to differ in cognitive demand on students. Instruction in U.S. schools has historically placed a high emphasis on procedural knowledge, the emphasis that is thought to be less cognitively demanding. I call this emphasis the common instructional approach. Research suggests that a more balanced and intertwined emphasis on procedural and conceptual knowledge, referred to as the blended instructional approach, will better support students' learning. Other research suggests emphasis depends on who is being

taught. Given new analytic methodologies, I investigated these theories using data from a daily teacher log and second through fifth grade student achievement measures collected by the Study of Instructional Improvement. I used four log items to define the instructional approach used in classrooms. Further, I use existing items and an IRT 2-parameter model to measure student's knowledge of multi-digit computation. Limitations for linking levels without linking items or linking group were unsolved. I applied hierarchical linear models, Rubin's causal framework, and propensity score causal inference techniques for studying causation. I found very few covariates systematically predict who receives the approaches. In the lower grades, school characteristics influence the instructional approach in use, but class characteristics influence the approach in use in the upper grades. From the causal analysis, students in classes receiving the blended instructional approach achieved the same as students in classes receiving the common instructional approach. Overall, this investigation found no support for the instructional approaches supported by the *Standards*. Furthermore, regarding analytic methods, this research concluded that future investigations comparing instructional treatments might benefit from using statistical methods that model treatments as they are received by students within and across academic years.

Chapter I

Introduction

This dissertation compares two approaches commonly used to teach multi-digit computation in elementary grades. The two instructional approaches under investigation differ in their emphasis on aspects of mathematical proficiency, often referred to as *procedural* and *conceptual* knowledge. The question—of which emphasis should be preferred— has been the center of debates in mathematics instruction for over 20 years. Instruction that emphasizes conceptual knowledge is thought to be more cognitively demanding, emphasizing student work on both standard algorithms and invented procedures, intertwined with justification and connections with place value concepts. Instruction that emphasizes procedural knowledge is thought to be less cognitively demanding. It places a very high emphasis on repetition of standard algorithms and little or no emphasis on invented or alternative procedures focused on justification and connections with multi-digit concepts.

Two Intersecting Problems

This dissertation addresses two problems, one related to instruction on multi-digit computation and the other connected to methods of measurement and statistical models. First, research on enacted mathematics instruction, in general, and instruction on multi-digit computation, in particular, has found that conventional instruction regularly places a high emphasis on procedural work. In contrast, research on mathematical learning suggests that instruction that intertwines procedural and conceptual knowledge enhances students' development of higher-order thinking skills and mathematical proficiency (Rittle-Johnson, Siegler, & Alibali, 2001; Stigler & Hiebert, 1999) Furthermore, professional organizations recommend that instruction incorporate a more balanced emphasis on procedural and conceptual knowledge (National Council of Teachers of Mathematics, 1989, 2000). Although this recommendation has appeal, there is no scientific evidence supporting the superiority of this balanced approach. Moreover, there is persistent evidence from national studies of student achievement that reveals deficiencies in student learning and achievement gaps between dominant and marginalized groups (National Center for Education Statistics, 2009, 2011). Research that seeks to investigate these gaps and deficiencies has found that students with low prior achievement are more likely than their counterparts to receive instruction that emphasizes procedures while their higher achieving peers are more likely

to receive instruction emphasizing procedures and concepts (Gamoran, 2010). Is one approach more effective than the other, and does the answer differ by student characteristics? Research investigating the effectiveness of the two instructional approaches has not directly attended to issues of varying effectiveness on group characteristics. The general position of many mathematics educators has been that, for *all* students, development of higher-order thinking skills is more well supported by instruction emphasizing both procedures and concepts.

There are other reasons to be concerned about the classroom use and effects of the two instructional approaches under investigation. Findings from research on teacher selection and implementation of instructional programs have identified relationships between ethnicity, social class, and socioeconomic levels and "slow-paced, dead-end" instruction (Gamoran, 2010, 2011; Oakes, Gamoran, & Page, 1992). Related to this investigation, these findings suggest that minority students from low income families are more likely to receive instruction that emphases procedures and less likely to support development of higher-order thinking skills. Some educators argue that difference in use of instructional approaches may be needed and reasonable. Their argument rests on claims that teachers may be responding to students' developmental needs constrained by instructing large classes and not their minority status (Ansalone & Biafora, 2004). Are differences in emphasis on procedures and concepts responsive to students' needs or

the result of social factors? From teachers' professional judgment, there is evidence that more than one approach may be valuable and productive. Early research agreed with this observation (Peterson, 1979).[i] Since teachers work in isolation from their peers and they have autonomy over class endeavors, instructional decisions such as these are often rooted in personal or idiosyncratic warrants (Lortie, 1975).

There are reasons for the persistence of this fundamental controversy. Research evidence is insufficient to guide instructional polices. The reasons for this lack of evidence are complex. Overall, these instructional issues have not been easy to investigate, despite ongoing interest and will among educational researchers to investigate them. Three challenges stand out.

First, fundamental measurement issues have plagued this line of research. Specifically, research has lacked systematic methods for recording instruction in detail and at scale. Researchers primarily focused on teaching practices, ranging narrowly from planning to classroom activities, with little attention to instructional approach or content (Rosenshine & Furst, 1973; Shavelson, Webb, & Burstein, 1986). Records of practice relied on frequencies recorded using Likert type scales, categories and tally marks without pertinent connections or teacher input. Furthermore, data have been collected mainly in classrooms without attention to factors outside the classroom —for example, school climate on academic performance or pressures on

teachers to use innovative instructional approaches—that might influence classroom activities. Samples were mainly small, ranging between 20 to 100, due to observer time constraints and the cost of hiring large numbers of observers (Good & Grouws, 1977; Wallen & Travers, 1963). Concerns over inter-rater reliability and bias due to an inadequate number of variables have been raised.

More recent studies have used teacher logs or diaries. Data collected retrospectively and gleaned from early versions of these instruments often suffered from problems of memory distortion and inaccuracy that arose when respondents were asked to summarize behaviors they engaged in over an extended period (Bradburn, Sudman, & Wansink, 2004; Hilton, 1989; Hoppe et al., 2000; Leigh, Gillmore, & Morrison, 1998; Lemmens, Knibbe, & Tan, 1988; Lemmens, Tan, & Knibbe, 1992; Sudman & Bradburn, 1982). However, newer teacher log designs and log items have been shown to produce reliable data on instruction and with greater depth on core content of the elementary grade than by previous means (Camburn & Barnes, 2004). However, using these new teacher records for identifying the instructional approach used in the classrooms during instruction on a particular content has not been explored.

Second, this research often lacks meaningful student achievement measures that mirror the content being taught. When student achievement is the outcome of interest, researchers rely heavily on

measurements of knowledge on subject basic skills using standardized tests. Typically, test items and students' responses are not available to researchers. This is generally the situation today, but greater accessibility to computers, enhancements of computational power, and psychometric software have changed the outlook for the use of test items and the analysis of outcomes. Research, however, has not explored these possibilities.

Third, this line of research also lacks advanced statistical methods that accurately model the instructional environment, control for mediating effects, and produce answers backed with scientific evidence. Unfortunately, researchers rely heavily on an exploratory disposition, using descriptive statistics, correlations, and, too rarely, analysis of variance. In the end, these methods only produce descriptions of central tendencies and bivariate relationship without control for confounding variables that bias results (Stone, 1993). Furthermore, statistical methods available around the 1970s and 1980s limited researchers' ability to investigate the varying effects of instruction for different units, such as classrooms and schools, however, advances in statistical methods, particularly hierarchical linear models (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002), have expanded knowledge about how schools and classes affect student achievement. Recent research by Rowan, Harrison, & Hayes (2004) provides a good example. These researchers found that there is more variability in instructional practices

among classrooms within schools than between schools. This finding is at variance with prior models on school, classroom, and instructional effects. Presently, research now recognizes hierarchical models as the appropriate methodological approach for exploring instructional effects and attending to the nested school structure and the random effects for schools and classrooms.

Still, other advances in statistical methods have created new opportunities for readdressing the past theories on teaching effectiveness. Much of the past and present theories and inquiries on instructional effects on student achievement are about causal effects, yet researchers lack the appropriate statistical methodologies to provide essential answers. Randomized experiments are preferred for answering causal questions, as randomization ensures that pre-existing characteristics of experimental units are unrelated to treatment group assignments and statistical inferences on data about group differences provide unbiased estimates of the causal effects of interest. Unfortunately, random assessment is often not feasible (Cook, 2002). In such cases, methodologists have found Rubin's casual framework in conjunction with propensity scores to be useful (Imai & van Dyk, 2004; Rosenbaum & Rubin, 1983; Rubin, 1997). I discuss these methods in Chapter III. In general, results from studies using these methods provide better support for causal claims. Propensity score methods allow the researcher to match cases on probability of being treated, and, when

used in a linear model, test for treatment effects as if an experimental design had been used. Furthermore, these techniques are currently being adapted to multi-level or nested data, and there is sufficient support for use of this method in studying instructional effects on student achievement. Still, few studies have used these methods to investigate effects on student achievement of instructional approaches and establish causal evidence.

This dissertation sits at the nexus of these two problems – one being about quality of instruction and outcomes for students and the other related to how research can contribute to improvements. This dissertation asks the first question by developing ways to answer the question with rigor. To address the first question, I use a set of multi-level models, where two treatments are defined by differences in instructional approaches on procedures and concepts. In the first set of models, I investigate class, teacher and school characteristics that predict instructional treatment. Here, instructional approach is the dependent variable in two-level hierarchical models. For the class characteristics, I consider class ethnic composition, gender, initial achievement, and socioeconomic status. I also consider teachers' years of experience and several variables measuring professional preparations. At the school level, I test the school environment using averages on these measures. In a second set of models, I study instructional treatment effects on student achievement, the outcome, using a three-level model.

At the student level or level-1, I control for group characteristics in terms of prior knowledge. At the classroom level or level-2, this is where I control for treatment effect and propensity for being treated. This model also includes a school level at level-3. I define these models and covariates in more detail in Chapter III.

Data

This research uses data from The Study of Instructional Improvement (SII), originally a large-scale quasi-experimental investigation of the design, implementation and instructional effectiveness of three of America's widely disseminated Comprehensive School Reform (CSR) programs. This research relies primarily on daily teacher records of instructional practice and bi-annual student achievement assessments. The details of SII and its data are described in Chapter III. Compared to past teaching effectiveness studies, this research focuses on all elementary grades, sampled large urban populations, and uses advanced measurement techniques.

Instruction Emphasizing Procedural and Conceptual Knowledge

What is procedural and conceptual knowledge? For this dissertation, I utilize the widely established definitions of procedural and conceptual knowledge presented by Hiebert and Lefevre (1986) over 20 years ago.[ii] They define procedural knowledge as the familiarity with

mathematical symbols, their appropriate representation and use *and* to the rules or procedures for solving mathematical problems, and conceptual knowledge refers to knowledge rich in relationships and connections between pieces of information. In terms of work on multi-digit computation, procedural knowledge is the familiarity with symbols associated with addition, subtraction, multiplication, and division or specifically, the symbols +, -, x, and ÷. It includes standard representations such as

(a) $21 + 34 = 55$ or $55 - 34 = 21$

(b)
$$\begin{array}{r} 21 \\ +\ 34 \\ \hline 55 \end{array} \quad \text{or} \quad \begin{array}{r} 55 \\ -\ 34 \\ \hline 21 \end{array}$$

(c) $12 \times 12 = 144$ or $144 \div 12 = 12$

(d)
$$\begin{array}{r} 12 \\ \times\ 12 \\ \hline 144 \end{array}$$

(e) $\sqrt{144} = 12$.

Students who have multi-digit computation procedural knowledge are also able to carry and borrow and recognize contextual clues for when to use each operation and the appropriate representation such as the sum for addition and difference for subtraction. Alternatively, learners who recognize the relationships between addition and subtraction, multiplication and division, and addition and multiplication,

and use the opposite operation to check their work and identify mistakes have multi-digit conceptual knowledge. They understand the usefulness of the representations in (a) – (e), their knowledge of multi-digit number concepts carries over such that they understand what the numerals mean (their positional value or meaning) in each of the representations, and they know how to expand the representations for larger numbers. Additionally, they can develop alternative algorithms and make sense of alternative algorithms developed by others.

Relation to Instruction

In the past, mathematics curricula emphasized procedural skills (Hiebert & Lefevre, 1986). Research has found that instructional approaches emphasizing procedural skills spend a lot of time on executing computation algorithms accurately without attention to why — the multi-digit concepts that justify the steps of the algorithm (Mitchell, Hawkins, Jakwerth, Stancavage, & Dossey, 1999; National Research Council, 2001; Stigler & Hiebert, 1999). In these classrooms, instruction attends to dissecting algorithms into their sequential steps and language is on columns and not on place-value. Students in the class may individually or together practice over and over executing the steps of an algorithm. Seat work and home work entail more problems using the same skills and practicing the same series of steps. Manipulatives and

calculators are distractions as they draw attention away from the algorithm towards a focus on reasoning.

Researchers found that students who rely on only procedural knowledge can have "buggy" algorithms and misunderstandings and lack knowledge to repair their "bugs" (Brown & VanLehn, 1982). As a result, instructional guidance has pushed for more instructional emphasis on concepts and a de-emphasis on rote procedural skills (National Council of Teachers of Mathematics, 1989, 2000; National Research Council, 2001). Children who have multi-digit computation conceptual knowledge understand the relationships and connections among facts and procedures. For example, children can develop conceptual knowledge when instruction incorporates alternative algorithms and attends to why conventional and alterative algorithms work (Brownell & Moser, 1949). Instruction that attends to students' development of conceptual knowledge is still not the norm today (Rowan, et al., 2004).

Which of the two instructional approaches is most effective for students' learning of multi-digit procedural and conceptual knowledge has not been identified. How students develop procedural and conceptual knowledge and what is an appropriate instructional approach have been the center of debates for over twenty years. The current view is that the development of these two types of knowledge is intertwined and inseparable (Rittle-Johnson & Alibali, 1999; Rittle-Johnson & Siegler, 1998; Rittle-Johnson, et al., 2001). Still little is known about how

instruction should attend to each type of knowledge – to what order, emphasis, and duration. This has been essentially left for teachers to figure out.

Therefore, instructional approaches to teaching multi-digit computation vary greatly. Furthermore, within this variability, some approaches model practices at the center of the debates and follow the descriptions in the *Standards*. Still others differ substantially from these approaches. In this dissertation, I focus on the two instructional approaches at the center of the debates – one being primarily focused on procedural knowledge and the other blending an emphasis on procedural and conceptual knowledge. In the following section, I elaborate on the features of these two approaches.

Distinct Features of the Two Instructional Approaches

Instruction on multi-digit computation that places a strong emphasis on skill efficiency or procedural work has a long history in U.S. classrooms and is still common practice today. This practice defines one of the approaches examined by this inquiry. Research suggests that instruction that features this approach generally contains teacher-centered demonstrations, uses teacher-posed lower-order questions, is fast paced, and emphasizes error-free practice (Brophy & Good, 1986). Instruction that incorporates these features is thought to promote procedural competencies.

Conversely, our current theoretical understanding of how children develop procedural and conceptual knowledge defines the second approach examined in this dissertation. This instructional approach is thought to support children's development of both procedural and conceptual knowledge. In comparison to the first approach, this instructional approach places prominence on making important mathematical relationships explicit (Brownell & Moser, 1949; Fuson & Briars, 1990; Hiebert & Wearne, 1993). Research on this approach has found elements that cut across features of old dichotomies. For example, in some classrooms where this approach was used, teachers explained to the class the mathematical relationships using arithmetic procedures, a feature of the direct teaching approach. Still, in other instances, students developed their own algorithm and justified its validity, a feature of the inquiry-based teaching approach. Still in other classrooms, students were exploring relationships using symbolic presentations and concrete materials to compare similarities and differences – features that might align with the discovery, student-centered, and reform-based approaches. This second approach cuts across many of the salient features of "old" definitions of instructional approaches.

To define the two instructional approaches examined in this investigation, I focus on instructional work *with/on* algorithms. I use the features found in the definitions of what is means to have procedural and conceptual knowledge. These approaches do not place weight on who is

at the center of the work as in teacher-centered versus student centered instruction or direct instruction versus inquiry based instruction. I assume, for example, that if the teacher is demonstrating or discussing, then the student is also engaged in this work which is just as important as if the students were doing it themselves. I also do not attend to the pace of instruction or the level of cognitive demands associated with questioning. While practices share features with the two approaches under investigation, this information in not available in the SII data. Future research will need to define how to collect this data, whether by daily teacher logs or by other means.

Definitions of Common and Blended Instructional Approaches

The first of the two instructional approaches is similar in emphasis on skill efficiency as the one described in research as stated above. This approach is marked by attention to how to carry out the steps of a conventional procedure and students practicing doing so. This approach is called *common instructional approach* and is generally marked by a high instructional emphasis on procedures and little to no emphasis on concepts. The second theory holds that instruction should intertwine an emphasis on procedures and concepts. In this approach, instruction attends to both how and why conventional and alternative procedures work and to mathematical relationships. Students may also have opportunities for practice. I call this approach a *blended instructional*

approach. In contrast to a common instructional approach, it is marked by a high and more equal emphasis on procedures and concepts.

Research questions

I investigate the two problems using two sets of research questions. The first set of questions addresses the many parts of the problem related to instruction on multi-digit computation. The second set of questions addresses inquiries related to the use of methods of measurement and analysis while studying the two instructional approaches.

First research questions:

Does the blended instructional approach to teaching multi-digit computation to second through fifth graders cause greater learning for all students, as endorsed by the *Standards*?

Inside this first question, I investigate claims that led to or are related to this endorsement. In one investigation I study who gets the blended approach and what factors influence its use. Specifically, this investigation addresses whether teachers choose their instructional approach based on who they are instructing or by other factors. The other factors I investigate include teachers' knowledge, years of experience, professional preparation, gender, and ethnicity; curriculum material; class average ethnic, gender, and socioeconomic composition; and school environment. The second piece motivating this endorsement is related to causality. Here, I investigate whether classes who receive the

blended approach learn more than classes receiving the common instructional approach and if these gains in learning depend on the child's prior knowledge or their grade?

Second question:

Can new measurement and statistical methods be used to provide scientific evidence to an investigation on the causal effect of the blended instructional approach used in the teaching and learning of multi-digit computation?

Inside this investigation, I investigate the use of four items from a daily teacher log that measure differences in emphasis on multi-digit procedures and concepts to identify retrospectively classes receiving common and blended instructional approaches. Second, I investigate the use of items from standardized achievement tests, designed to measure general knowledge, to reliability measure student knowledge of multi-digit computation. Third, I investigate the usefulness of Rubin's Causal Framework and propensity score methods in assessing the causality of the blended instructional approach.

Organization of the Dissertation

The primary purposes of this dissertation are to test the use of new measurement and statistical methods that might permit/support answering long-standing questions about the use of instruction emphasizing multi-digit computation procedures and concepts in

elementary grades. Using an unusually large data set, I test the use of daily teacher log items to measure instructional approach, standardized test items to measure student knowledge of multi-digit computation, and multi-level models to study their relationship and occurrence.

I have organized this dissertation into five chapters. In Chapter II, I review the research on the teaching and learning of multi-digit computation. In Chapter III, I report on (a) measurement techniques used with the daily teacher log items, (b) psychometric methods used to measure students' knowledge on multi-digit computation, and (c) analytic methods, including Rubin's Causal Framework, propensity score methods, and hierarchical linear models. In Chapter IV, I report and summarize the results as they relate to the research questions. Also, in this fourth chapter, I discuss the limitations of this study. In the last chapter, I appraise the merits of the "new" measurement techniques and analytic methodologies used by this investigation, discuss the analytical results and related policy implication, and propose future research in terms of both studies on common and blended instruction and new inquiries on methods of measurements and analysis.

Chapter II

Instruction on Multi-digit Computation: A Review of Current Knowledge, Gaps and Opportunities for Future Research

Multi-digit computation is a broadly important mathematics topic. Competencies in this content area are needed for basic citizenship and for success in advanced mathematics topics including algebra. Therefore, effective teaching and learning of multi-digit computation is crucial. Fortunately, multi-digit computation is one of the most comprehensively studied topics of mathematics education, and given new instrumentation and methodologies, research within the topic is well positioned to identify well-founded ways to teach computational skills. Still, what *is* known about instruction on multi-digit computation? The purpose of this chapter is to appraise the nature and quality of knowledge about instruction on and learning of multi-digit computation with whole numbers in the U.S. as described in mathematics education literature since the 1940s. In this chapter, I pay particular attention to the state of evidence and position for defining instructional approaches and for whom the approaches are intended. But, before this discussion, I first

provide some definitions, review trends in student achievement, and delineate the scope of this review.

Defining Multi-digit Numbers, Computation Problems, and Curricular Significance

First, what are multi-digit numbers and multi-digit computation problems? For the purpose of this research, I consider a multi-digit number to be any whole number greater than 9. For example, 10, 101, and so on are multi-digit numbers. Furthermore, for this study, I restrict the set of multi-digit numbers to only positive integers greater than 9, leaving decimal representations of rational numbers for future research.

Multi-digit Number and Computation Problems

Multi-digit computation problem include computation problems involving addition, subtraction, multiplication, and division where the numbers involved have a total greater than 18. For example, 19 - 10 = 9 and 100 + 1 = 101 are included in the domain of multi-digit computation problems, and solutions to such problems as 21 ÷ 4 would be reported as 5 remainder 1. The decimal representation is outside the scope of this review. Furthermore, this study is not restricted to any one problem context – pure calculations problems, word problem,[iii] work with manipulatives, and using information from pictures or graphs – or models of addition, subtraction, multiplication, or division. For examples

of the different models, see Vest (1969, 1971); Bell, Greer, Grimison, and Mangan (1989); Fuson (1992b); and Greer (1992). Likewise, this review and research is not limited to particular algorithms; therefore, solutions to multi-digit computation problems may be obtained by standard or non-standard algorithms. Examples of these algorithms are given in Chapter III, Methods.

Curricular Emphasis

The priority placed on multi-digit number and computation is universal; in both the U.S. and other countries, it is a topic that students encounter early in their formal education and that continues throughout their elementary years. Furthermore, many children begin their formal schooling with informal computation skills and then in the elementary grades they develop formal knowledge of numerical operations – an understanding of and how to use whole-number computation algorithms (National Council of Teachers of Mathematics, 1989, 2000). In U.S. schools, there is a general curricular pattern in which children matriculate in learning computational skills, beginning first with an emphasis on whole number concepts and computation. In either kindergarten or first grade, children in begin with single-digit number concepts and computation when informal computational procedures are adequate and the essential conceptual knowledge centers on number — sequencing and cardinal meaning — and manipulation of objects. Then

later in these early grades, children begin to learn more formal computation procedures, most likely the standard single-digit computation procedures and notations and still base-ten concepts are not yet essential for success. In first and second grades, children begin work on multi-digit computation procedures and concepts; here, base-ten concepts become essential for children to become proficient.

There is also a general curricular pattern of exposure to the various operations. In the early grades, children focus primarily on addition and subtraction and begin to study multiplication and division, and then, while in grades three through five, they are expected to develop fluency in whole number operations (National Council of Teachers of Mathematics, 1989, 2000). This allocation of focus also approximates the emphasis reported by teachers. From the 1996 NAEP mathematics assessment, teachers of 93% of the fourth graders reported giving number and operations "a lot" of instructional emphasis (Allen, Carlson, & Zelenak, 1996). In a more recent investigation, Rowan, Harrison, and Hayes (2004) found that 39.5%, 40% and 41.9% of lessons in first, third, and fourth grades, respectively, included emphasis on multi-digit computation. Of these lessons, work with whole numbers comprised 91.8%, 82.4%, and 76.2% of lessons in first, third, and fourth grades, respectively. Therefore, multi-digit computation, within the standard U.S. curriculum, is a major part of the first- through fifth-grade curriculum.

Achievement Concerns Related to Multi-digit Number and Computation

Despite the likelihood that students have years to learn multi-digit concepts and computation procedures, there is still reason for concern about achievement in this content area. Educational researchers and policy makers know from national and international studies that elementary students' achievement is below desired levels.

Results from National Studies

Results reported for mathematics in 2005 of the National Assessment of Educational Progress (NAEP), an assessment comprising 40% of items on whole number concepts and operations, showed small gains over previous years' results (Perie, Grigg, & Dion, 2005). Only 36% of fourth graders performed at or above the proficient level. These results have been fairly consistent for many years with reform efforts resulting in little change in these results.

Specifically, results from more resent assessments have showed shown insignificant improvements and reports have emphasized these findings (National Center for Education Statistics, 2009, 2011). From 1990 to 2011, scores have only increased by 28 points, on average, over the 21 years of testing. Furthermore, for the 2007, 2009, and 2011 assessments, there has been no increase in the number of students reaching "basic" level. On all of these recent assessments, 82 percent of students reached "basic" level and were able to compute the difference

between two 4-digit numbers, for example; however, 18 percent of fourth graders were unable to solve such problems. (Note that 40 percent of fourth grade assessment items focused on number properties and operations.)

NAEP results also reveal differences in mathematics achievement between student groups represented by different demographics. For example, since 2003, NAEP has recorded a 25-point or more gap between African Americans and Caucasians and at least a 20-point difference between whites and Hispanics and Caucasians and American Indians. For the 2011 assessment, researchers compared the demographic profiles of students scoring below the 25th percentile and those scoring above the 75th percentile. Students in the lowest percentile compared to the highest percentile comprised 41 percent fewer Caucasians, 23 percent more African Americans, 24 percent more Hispanic, and 8 percent fewer Asians. They were also 51percent more eligible for free or reduced school lunch, 20 percent more with disabilities, and 19 percent more as English language learners. In sum, differences in mathematics achievement are strongly related to socio-economic differences.

International Comparison Studies

Results from the Third International Mathematics and Science Study (TIMSS) also underscore the need for improving K-5 student mathematics achievement. These results have called attention to the

elementary mathematics curriculum indicating that U.S. children's mathematics achievement for basic computational skills lags behind their Chinese, Japanese, and Korean peers in their computation skills (for a list of publications, see http://nces.ed.gov/nations reportcard/; http://nces.ed.gov/timss/). The TIMSS assessment given to 9 year olds (for most countries this was third and fourth grades) contained 102 items with 25% of these items assessing whole number competencies which tested a range of performance expectations — knowing, performing routine procedures, using complex procedures, and solving problems (Mullis et al., 1997). Compared to 17[iv] other countries, United States fourth grades mean mathematics scores were significantly lower than that of seven countries and significantly higher than that of 12 countries. Alternatively, Singapore and Korea had scores that were significantly higher than all other participating countries. In comparing results for third graders, the United States was ranked sixth among the 16[v] comparable countries. Overall, these results suggest that, on average, 9 years old in the United States can perform higher and become more proficient in primary-school mathematics.

The findings are consistent with reports on the lag in student achievement scores in mathematics and the persistent gap between student groups. Reasons for these findings are, however, less consistent and obvious. Therefore, given the longstanding poor mathematics achievement and persistent achievement gaps, there is a need to

understand what is known about the teaching and learning of multi-digit computation and to identify areas which have not been addressed in the research.

Delineating the Scope of this Review

This review focuses on research on the teaching of whole number multi-digit computation procedures and concepts to students in elementary grades in traditional U.S. school and classrooms. In the sections that follow, I provide my reasons for this limited focus.

Rationale for Focus on Whole Number Multi-digit Computation Procedures and Concepts

I focus this review and the research that follows in subsequent chapters on whole number multi-digit computation for several reasons. First and fundamentally, as previously stated, knowledge of multi-digit computation procedures and concepts is considered minimal knowledge needed for citizenship. Thus, knowing the state of knowledge gained from research on this content is essential to identifying ways of advancing the teaching and learning of this content and ensuring that all students gain this minimal knowledge.

Second, there is evidence that students' experiences with whole numbers is insufficient (Hiebert, 1992). The mathematics represented by whole numbers is less complex than the mathematics of decimals,

fractions and integers, still preparation for this more complex mathematics work begins with work on multi-digit whole numbers. Specifically, there is evidence that once children begin operations with integers and decimals with choice of operation, errors often arise from misconceptions (Bell, Fischbein, & Greer, 1984; Bell, Swan, & Taylor, 1981). Common misconceptions include multiplication always makes bigger, division always makes smaller, and division must be of a larger number by a smaller one. These misconceptions arise, in part, due to students' generalizations drawn from problems limited by range of type of number, linguistic structure, and context. Much of this research was conducted in the 1980s and 1990s. One key finding was that textbooks provide too little support and experience that students needed early on in order to make sense of later instruction. This finding supports looking closely at the teaching and learning of whole number multi-digit computation, since there is evidence suggesting there are weaknesses in students' learning experiences with whole numbers and in their preparation for work on advanced topics.

The third reason centers on the fact that the teaching and learning of whole number multi-digit computation has received considerable attention since the 1940s, and consequently instructional guidance from professional organizations and reforms, in general, have targeted this content area. Much of the guidance has been in effect for over 20 years. Furthermore, this evidence suggests that the reform practices are

warranted, but achievement still lags for many children on learning this very content area. Therefore, this narrow review looks closely at the state of knowledge of the very goals of a major line of research in mathematics education.

In sum, looking closely at the state of knowledge of the teaching and learning of whole number multi-digit computation is warranted for its core value for citizenship, its prerequisite position, and the continuous attention it receives by educational researchers and policymakers.

Rationale for Limiting Review to Studies Conducted in the U.S.

There are two primary reasons for limiting this review to studies conducted in the U.S. First, international comparison studies have revealed the importance of culture in understanding differences in teaching practices and students' learning (Stigler & Hiebert, 1998, 1999). Furthermore, instructional activities, like other cultural activities, are highly stable over time, and they are not easily changed. Since instructional activities are embedded in the wider culture and not readily apparent to its members, understanding differences in these activities from country to country will be difficult through deliberate study by non-members. Further, even when effective practices and approaches are identified, cultural activities do not export well into other countries.

Therefore, research conducted outside the U.S. is weakly informative to the research discussed in this dissertation.

Second, language plays an important role in learning multi-digit concepts, particularly in the early grades. Furthermore, in general, language plays an important role in instructional choices and student learning. For example, in the English language the base-ten system is not readily apparent in the number naming for numbers between 10 and 20 while in other languages the base ten system is apparent in the naming of numbers. Therefore, students learning the multi-digit number concepts for the first time using English as the primary language have more to learn and instruction must support their learning and account for the irregularity in the number naming and lack of support for the base ten system gained in many other languages. Given these fundamental language differences, research on instructional and students' approaches to multi-digit computation on whole numbers when the language in use is not English, in the end, yields inadequate information.

Therefore, I limit this review to primarily research conducted in the U.S. schools or on children receiving their education from schools located in the U.S. and when the primary language of instruction is English.[vi] For some research that has been excluded, I have provided references in the text which may be helpful to interested in readers.

Summary Timeline of Research

The development of a line of research that examines instruction on multi-digit computation and its impact on student learning began sometime prior to the 1940s. The earliest identified research within the mathematics literature occurred in 1949 by Brownell and Moser, however, these researchers acknowledge in their paper that the topic of their inquiry had engaged researchers for thirty or more years. So, while research on multi-digit computation began more than a century ago, interest in this topic gained momentum and rigor in the 1980s.

During the 1980s, research on multi-digit computation flourished, with most being observational in nature and focused on understanding how children approach computation, primarily with addition and subtraction. Much of the research ranged from studying the informal ways that children approach computation (Carpenter & Moser, 1984; Fuson, 1982, 1984; Siegler & Booth, 2004; Siegler & Shrager, 1984), children's difficulties and errors (Baroody, 1984; Brown & VanLehn, 1982; Fuson, 1984; Hatano, 1982; VanLehn, 1986), instructional approaches (Lampert, 1986), and instructional effects on achievement (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989). Included in this decade is also one study on the comparison of the U.S. elementary mathematics curriculum with several non-U.S. curricula (Fuson, Stigler, & Bartsch, 1988).

Research in the early 1990s on multi-digit computation came in support of and in response to the 1989 NCTM *Curriculum and Evaluation Standards for School Mathematics* (aka. *Standards).* The *Standards* are just one the many attempts in recent times to organize instructional guidance, and reform teaching, and improve learning. Of particular interested to this dissertation is the guidance provided in the *Standards* which reference instructional practices on multi-digit computation. Here, the *Standards* These recommends suggest a decreased emphasis on mechanical drill and memorization and an increased emphasis on mathematical reasoning and understanding, but, at the time of its publication, justification for these recommendations was dismal at best. Research, since the introduction of the *Standards*, has focused on providing evidence for the recommendations and thus has focused on the teaching and learning of multi-digit computation with emphasis on mathematical reasoning and understanding and on mathematical content knowledge needed for teaching.

For the purpose of this review, I focus on research conducted since the 1940s and place more attention on research conducted from 1970s to present.

Topics Reviewed

To synthesize research relevant to instruction on multi-digit computation,[vii] I identified five primary research topic areas discussed in

the research literature. The first two areas of concern are the critical components of student learning: the mathematics to be learned by elementary school students and students' approaches to this mathematics. Another area of interest is instruction and student learning, namely instructional approaches that benefit students' learning in this domain. The final two areas of relevance are the critical components of teacher knowledge and learning: the knowledge teachers need for teaching and opportunities for learning to teach in this domain.

In order to highlight relevant knowledge and identify gaps in the research and opportunities for future research, I have divided this review into two sections. In the first section, I discuss knowledge as it is comprised in five topic areas of the larger domain of instruction on multi-digit computation. Specifically, the five topic areas are (1) the mathematics to be learned by elementary school students, (2) students' approaches to this mathematics, (3) instructional approaches that benefit students' learning in this domain, (4) the knowledge teachers need for teaching in this domain, and (5) teachers' opportunities for learning to teach in this domain**.** In the second section, I summarize the prevailing and identified gaps in knowledge on multi-digit computation as they occur across the areas.

Discussion of the Five Topic Areas Covered by Research on Multi-digit Computation

Areas Covered by the Research

Research covering the five topic areas generated knowledge generated from a century of investigations into multi-digit computational concepts. By considering the research specific to each of these five areas, additional themes for research and opportunities for future examination emerge. For each area, I discuss the prevailing findings and the areas in need of additional investigations.

Generally, there is unequal treatment of the operations, grades, and areas across the research. Specifically, there has been more research on addition and subtraction, and, for the most part, the research has focused on addressing areas 2 and 3 - " the mathematics students need to learn" and "the instructional treatments that are effective in helping students learn the mathematics."

As I discuss the findings and evidence, I pay particular attention to the methods used and type of evidence on which findings rest. There is a present interest in knowledge gleaned from scientific studies. Therefore, as part of this review, I appraise the expanse of knowledge from such studies.

The Mathematics Content or Concepts that Students Need to Learn

Findings. Ten research articles directly address the mathematics content and concepts that students need to learn related to multi-digit computation. Eight comprehensive discussion or review articles address multi-digit content and conceptual knowledge needed and constructed by children. First, this line of research has identified the range of real world situations modeled by addition, subtraction, multiplication, and division which in turn necessitate learning the meaning of number, the different meanings for +, -, and =, how to carry out the different operation, and how to communicate and use mathematical ideas (Fuson, 1992b; Greer, 1992). The three additive situations include "change add to," "combine physically," and "combine conceptually." Similarly, there are three subtraction situations – "change take from," equalize," and "compare." Fuson (1992b) provides examples of each of these and specific examples of word problems involving whole number addition and subtraction. Vest (1969) gives an even broader range of examples. Multiplication and division derive from a somewhat different set of real world situation. Research has summarized the situations into 10 classes – "equal groups," "equal measures," "rate," "measure conversion," "multiplicative comparison," "part/whole," "multiplicative change," "Cartesian product," "rectangular area," and "product of measures" - which are not entirely distinct from each (Greer, 1992). In fact, in some case, it is the person solving the problem who determines the class in which the problem

belongs and its related approach. Greer gives examples for each class for both multiplication and division. Anghileri (1989), Kouba (1989), and Vest (1971) also present a range of situations modeled by whole number multiplication and division with some examples specific to use with manipulatives.

Researchers have studied the conceptual knowledge children build for competencies at numbers and addition, subtraction, multiplication, and division with numbers up to 100 and, somewhat separately, the multiunit numbers greater than 100. Sensibly, much of the knowledge children build during work with numbers up to 100 overlaps with the conceptual knowledge needed for work with larger numbers.

Research on unitary conceptual structures has documented the knowledge sequence that young children build starting with the spoken number names in sequence (Fuson, Richards, & Briars, 1982). For example, many children begin counting with knowledge of the number sequence that is undifferentiated such as "onetwothreefourfivesixseven." Furthermore, Fuson and her colleagues have mapped out children's developmental sequence for the count and cardinal meaning for number[viii] and number names with the last developmental level being "bidirectional chain/truly numerical counting." At this final conceptual level, children's knowledge of number is flexible. They understand a number word is a sequence word, with sequential words one apart, and a cardinal word, where the *n*th ordinal number word is preceded by a

cardinal group of $n - 1$ objects. Knowledge at this final level is key for computation. Specifically for addition, children now see the two addends as separate quantities from the total and addends can be partitioned to construct different addends. For example, in problems such as 7 + 5 = ? Children can now re-write 7 as 5 + 2 and use their knowledge that 5 + 5 =10 and 10 + 2 = 12.

A second related conceptual area addressed by this line of research includes English number words (i.e. six, eighty-four), written marks (i.e. 6, 84), and positional value (i.e. 84 represents 8 tens and 4 ones) that children construct to interpret and solve whole number addition, subtraction, multiplication, and division situations (Fuson, 1990a, 1990b, 1992b, 2003; Fuson, Fraivillig, & Burghardt, 1992; Fuson & Kwon, 1992; Fuson, et al., 1982; Fuson et al., 1997). This research suggests that children need to understand the features of the English number words and the system of written number marks (including positional values) and how the different multi-digit operations are related to each of these systems. Specifically, there are two conceptual structures for both number marks and number words that children need to learn. For number marks, children need to learn the visual layout and that the positions in number marks are ordered in increasing value from the right while spoken, and then for written number words, they need to learn the unit names for each position and the decreasing order in which they are said. Two additional conceptual structures support multi-digit

addition and subtraction. Children must learn both the quantities that each multiunit represents and the ten-for-one and one-for-ten trades between the neighboring positions.

Research has studied the relationship between the system of number marks and the system of English number words and their support of multi-digit addition and subtraction, in particular (Fuson, 1990a, 1992a, 1992b; Fuson & Smith, 1997; Fuson, Wearne, et al., 1997). This research has found that these systems contribute different aspects to multi-digit addition and subtraction. Specifically, when adding or subtracting, the quantities of each multiunit represented by the English number words direct the operation of like multi-units. Further, this research found that the English number words can direct and constrain correct trade rules when one has too many or not enough of a given multiunit quantity and that the written marks require trading in multi-digit addition and subtraction because each multiunit is limited to 9 or less.

Children learn the number names, written marks and positional value but not without some obstacles. A prominent line of research has studied the irregularity of English number names and children's conceptual understanding (Fuson, 1990a; Fuson, et al., 1992; Fuson & Kwon, 1992; Fuson, et al., 1982; Fuson & Smith, 1997; Fuson, Wearne, et al., 1997). Through observational analysis, Fuson, either alone or with colleagues, documented the irregularity in number names in the system

of English number words (in the teen words, decade prefixes, and decade names), in the English number named-multiunit system, in children's conceptual understanding, and in children's multi-digit addition and subtraction errors or misunderstandings. Of specific issue are the spoken words for the numbers between ten and one hundred. The irregularities include special words (i.e. twelve), different pronunciations (i.e. four, fourteen, and forty), reversals in the teen words (i.e. nineteen), and two different modifications of "ten," "teen" and "ty." The system of written marks, however, is a regular relative positional system. The discussion papers just mentioned discuss this issue, speculating that the irregularities in the English system of number words for two-digit numbers induce children to use unitary and concatenated single-digit conceptual structures rather than multiunit conceptual structures for multi-digit numbers (Fuson, 1990a). Fuson and Kwon (1992) expose this issue more concretely in a study involving only Korean second and third graders. Here, they reported on these children's understanding of place value and multi-digit addition and subtraction as reflected by their solution procedures and errors on multi-digit problems and by the trading involved in such problem. These researchers found at least 93% accuracy rate in computations involving addition and subtraction with trading of 2- and 3-digit numbers, with the exception of second-graders work on 3-digit subtraction with trading which had an accuracy rate of 73%. Fuson and Kwon argue that spoken language, written symbols, and

aspects of culture can support or interfere with children's construction of conceptual structures that facilitate learning 2- and 3-digit addition and subtraction. Prior to the research by Fuson and her colleagues, many researchers had been trying to hypothesize what may account for the apparent facility by which Asian children execute procedures. Fuson et. al. highlighted the invisible—the language used to name numerals and to operate numerals completely exposes decimal place value in Asian languages and that language actually shapes thought.

A longitudinal study of Chinese and U.S. 2-, 3-, and 4- year olds provides some evidence of the affect of the irregularity in the English number words (Miller, Kelly, & Zhou, 2005). Using survival analysis, these same researchers found no significant differences between the percentage of preschoolers who could count to ten, however counting from 10 to 20 is significantly easier for Chinese children. Then, of the children who could count to 20, there was no significant difference in the percentage of U.S. or Chinese preschoolers who could count to 100.

Research studying the conceptual structures needed for multi-digit numbers has identified four advanced concepts that support multi-digit multiplication and division which are related to trades and the base-ten structure of the English number system (Fuson, 2003). Two concepts refer to the cumulative relationship between positions and values. For one, children need to learn the relationship between the number of trades and the position. For another, they need to understand the

relationship between the position (or value) and multiplies of 10. For example, the fourth position or the thousands position is achieved by three trades and is three multiples of ten. The last two conceptual structures are related to exponential notation for multiplies of ten, one in words and the other in number marks (i.e. “ten to the third power” and 10^3).

In other earlier research, Lampert (1986) studied conceptual knowledge needed for multi-digit multiplication and division. She described the principled knowledge beneficial for multi-digit multiplication includes knowledge of place-value, additive composition of numbers, associativity property of addition, commutativity of addition, multiplicative composition of numbers, associativity of multiplication, commutativity of multiplication, and the distributive property of multiplication over addition. She also claimed that the mathematical knowledge required for multi-digit division includes place-value concepts and a conceptual understanding of ratio and proportion (Lampert, 1992).

Weaknesses and gaps. Research addressing the conceptual knowledge needed for work with multi-digit numbers and multi-digit operations has covered the span of content children experience. However, there are opportunities to learn more about the knowledge children need to learn related to multi-digit computation.

First, the research makes convincing arguments about the 10 conceptual structures of multi-digit numbers and their usefulness to

multi-digit operations. This research, however, does not attend to *how* children build these conceptual structures. In comparison to the detail given about how children build conceptual structures of counting and the sequential and cardinal meanings of numbers, the order and minute steps in knowledge development, the known conceptual structures related to multi-digit numbers and operations are macro or highest/end concepts. There are likely intermediate conceptual structures that children build in the process. Researchers acknowledge that there may even be additional levels in the conceptual structure. Still, research has not attended to this knowledge.

Related, research has not specifically attended to order in which the conceptual structures of multi-digit number develop. I can imagine that some concepts develop concurrently, yet other concepts require prerequisite knowledge before being learned. This kind of knowledge about the relationship between conceptual structures has received little attention.

These shortcomings are due to the limited range in research approaches used in this line of research. Except for the research by Lampert on fourth graders, this research has focused mainly on children between the ages of 2 to 8 years old. Furthermore, more research on multi-digit number and operations is based on multi-digit numbers less than 100 than on numbers over 100. Both the focus on numbers less than 100 and on "young" children have contributed the shortage of

knowledge on multi-digit multiplication and division. There are good reasons for studying older elementary age children and their conceptual development of multi-digit number and operations using larger numbers. First, multi-digit addition and subtraction is taught for several years beginning in second grade and continuing for several years. Second, children's work on multi-digit multiplication and division usually follows work on multi-digit multiplication and division and continues possibly into middle school grades. Third, reports on national assessments suggest that a significant proportion of children struggle to learn this content. Therefore, there is likely more to learn about the conceptual structures children older than 8 years build and attending to this limitation might alleviate deficiencies in achievement.

How Students Approach Multi-digit Computation

Findings. I identified 22 articles that contribute knowledge on children's approaches to multi-digit computation. Approximately 14 comprehensive studies or discussion articles address addition, 11 articles address subtraction, four articles address multiplication, and two articles address division. Some of the articles dealt with more than one operation, usually addition and subtraction were coupled in one study and occasionally addition and multiplication were coupled together. One study involved 1400 students or 41 classes and another involved 336 students, but typical samples ranged from 1 to 6 classes or 1 to 132

students. Ten studies involved students that ranged from first through third grade. One study used third graders only and still another study used third, fourth, and fifth graders. One study included first through fifth graders and three studies were multi-grade but not described by the other "grade" categories. Twelve studies were interpretive or observational in design, while the remaining five studies were comparative.

Children's approaches to addition and subtraction cluster under three general categories: strategy use (includes manipulatives, invented algorithms, achievements or lack there of (errors)), errors, and procedural supports or "crutches."

As one would expect, children's computational strategy choice or use becomes more mathematical over time. Children begin to solve computational problems by first modeling the problem as stated and moving onto partial modeling and then to strategies that involve no modeling. The final stage of strategies involves number facts. For example, in one study, a longitudinal study of first through third graders, children's addition strategies were found to begin with counting-all, then move onto counting-on, and then to include number facts (Carpenter & Moser, 1984; Fuson, Wearne, et al., 1997; McIntosh, 1998).[ix] Children initially count-on by ones but eventually learn to count-on by tens and ones (Fuson, 1982). Modeling and counting strategies are used before formal instruction on arithmetic begins and used after several years of

formal instruction (also see Ebby (2005); Fuson (1982); Fuson and Willis (1988)). In this same study, children's strategies for joining missing addend problems involved modeling and counting that reflected the additive action of the problem. The one variation on the pattern of problem solving strategies entailed subtraction for separate problems. Children initially modeled the problem using the separate-from strategy and this strategy gives way to using number facts, not counting strategies. In another study of third, fourth, and fifth-graders' multiplication strategies, students were found to begin with direct modeling and as their conceptual understanding of multiplication develops, they begin to use complete number strategies followed by partitioning strategies (Baek, 1998). Furthermore in a cross-sectional observational study of 33 second-grade students of low, middle, and high achievement, children were presented during individual interviews with one single-digit subtraction problem and five multi-digit addition or subtraction problems (four were addition and one was subtraction) (W. M. Carroll, 1996). Of the multi-digit problems, two word problems and three numerical problems were presented in horizontal format. Students' responses to these problems were categorized as mental, standard written algorithm, counting by ones, and skips (not attempted or not complete). Researchers summarized response categories: 48% of the time students solved a problem mentally, 35% of the time students used the standard written algorithm, and 5% of the time students used a more

error-prone counting-by-ones strategy. Counting-by-ones lead to an error half the time. Many of these students showed some flexibility in choosing a procedure. For example, students who had used a mental procedure for solving two-digit problem switched to the standard written algorithm on the three-digit problem. Additionally, these children used a number of different mental procedures.

The observational nature of this study allowed researchers to study children's strategy choices but the effect of instruction on strategy choice was not apparent (W. M. Carroll, 1996). The student population was chosen from three school districts, and it is not clear whether students attending the same also attended the same class or different classes. Since instruction varies among schools and among classrooms within schools (Rowan, et al., 2004), there is reason to believe that students' strategy choices are affected by the instruction they receive. Therefore, the variation in student's strategy choices found in this study may be in affect a measure of variation in instruction – not children's strategy choices.

In a study intended to identify gender differences in strategy use, no significant gender differences were found for correctness and strategy use for number facts, and addition, subtraction, and non-routine problems across the three years of study (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998). However, when the researchers analyzed the data by year, they found strong and consistent gender differences in the

strategies used to solve multi-digit addition and subtraction word problems and computational tasks. Girls tended to use more concrete strategies like modeling and counting, while boys used more abstract strategies that reflected conceptual understanding. Additionally, at the end of third grade, the girls used more standard algorithms than the boys, and boys were more successful at extending procedures flexibly. The ability to solve extension problems in third grade appeared related to the use of invented algorithms in earlier grades.

There is evidence that problem presentation and presence or lack of presence of manipulatives affects children's use of strategies (Fuson, 1982; Fuson, et al., 1992). The use of manipulatives with multi-units helps children construct the multiunit-names, multiunit-quantities, and regular ten-for-one and one-for-ten trades conceptual structures. Base-ten blocks were found to help but coins did not (Fuson & Briars, 1990). There was considerable variability in how children drew the ten-sticks and dots, enclosed 10 dots, and showed the answer (Fuson, Smith, & Lo Cicero, 1997).

In early multi-digit computation, children must differentiate the tens from the ones. To facilitate this differentiation, children invent varied and elaborate scaffoldings to mark which are tens and which are ones. They underlined tens, drew loops and lines to connect the tens, drew separating lines between tens and ones, and labeled them. If children have to spend much attention to what goes with what, they can

easily get memory overload, lose track of what they are doing, and forget either the numbers involved in the situation or their already obtained partial results.

It is not surprising that children approach multi-digit addition by adding the numbers from left to right. They are, of course, engaged in learning to read which proceeds left to right. In a study of first- through fourth-graders' multi-digit addition strategies where some classes of students were being taught algorithms and some were not, children in the "no algorithms" classes worked on the addition problems (similar to 6 + 53 + 185) from left to right, adding the tens, then adding the ones, followed by adding the two sums (Kamii & Dominick, 1997, 1998). Children in the "algorithms" classes, as expected, worked the problem from right to left using the standard addition algorithm. Further, the "no algorithms" classes, both in the second and the third grade, produced the highest percentage of correct answers (45 and 50%, respectively). The "no algorithms" second- and third-grade classes produced more correct answers than all the fourth-grade classes, who were all taught algorithms. The incorrect answers of the second- and third-graders in the "no algorithms" classes were more reasonable than the wrong answers of the "algorithms" classes.

A common error occurs when children have inadequate valueless conceptual structures for multi-digit numbers, termed concatenated single-digit structure (Fuson, 1990a; Fuson & Smith, 1997; Fuson,

Wearne, et al., 1997). Others have identified more than 26 impoverished subtraction procedures (Brown & VanLehn, 1982). These researchers believe that children's errors in procedures are systematic with some instabilities.

"Crutches" that facilitate learning are easy for children to learn and use and can be easily discarded once a child no longer needs them (Brownell & Moser, 1949). Here the crutches children used arose in the form of borrowing notations that supplemented the language patterns with perceptual patterns and showed concretely and visually the rationale of the procedure. These crutches were taught to the children, not self invented crutches. Carpenter and Moser (1984), on the other hand, found in a study of 88 first through third graders that children develop their own strategies that make the procedure more concrete and visual. They found that children are able to solve addition and subtraction problems using a variety of modeling and counting strategies even before they received formal instruction in arithmetic. Similar results were found by Fuson and Willis (1988) in a study of children's use of counting-up to solve subtraction problems. Here, first and second grade children who were taught counting-up with finger patterns could use it to subtract numbers with as many as 10 places. Additionally, they claim that children can be expected to invent counting-up strategies after work on counting-on strategies and before instruction on counting-up strategies. Most of these invented strategies evolve from work on compare

and equalize subtraction word problems. Further, students have demonstrated significantly fewer systematic errors when allowed to invent their own problem solving strategies (Carpenter, Franke, Jacobs, Fennema, & Empson, 1998).

Weaknesses and gaps. Much of the knowledge on children's approach to multi-digit computation comes from studies of similar design - studies of mainly first and second graders who were presented some addition and/or subtraction problems during individual interviews. Children's approaches described here occurred under atypical settings, an interview setting and not during instruction from their regular teacher. What is not clear from this research is how children's approach is affect by context, problem choice, and instruction.

Many researchers advocate for instruction that accounts for how children approach multi-digit computation. Yet, research does not provide knowledge on children's approach to multi-digit computation in typical instructional settings, usually thought to be instruction given by their regular teacher. Furthermore, instruction varies between classrooms within schools and between schools. The observational nature of most of the studies on children's approaches to multi-digit computation does not allow researchers to account for the variations in instruction that students receive. Without controlling for instruction, it is unclear whether the variation in strategy choices and approaches to

computation is due to children's choices or teachers' instructional choices.

Studies varied in problem choices used to assess children's approaches. The range of problems included computational tasks, word problems, addition, subtraction, multiplication, division, two-digit numbers, three-digit numbers, and so on, but no studied included all problems. Each study's findings are limited but the problems used, as it is expected that students' approaches are affected by the problems presented to them.

Additionally, research needs to occur on upper elementary children and on multiplication and division. Presently, knowledge on children's approach to multi-digit multiplication and division is nearly non-existent.

Instructional Treatments that Are Effective in Helping Students Learn Multi-digit Computation

In this section, I focus my review on instructional treatments or practices that are specifically and uniquely effective in helping students learn multi-digit computation skills and concepts. Therefore, I exclude practices and treatments that are more generally effective in helping students learn mathematical skills and concepts. For example, Anghileri (2006) recently reported on scaffolding practices that enhanced mathematics learning. Such practices maybe important components of effective instructional approaches. However, since scaffolding practices

and similar practices are not unique to instruction on multi-digit computation, they are not reviewed here.

Findings. Nearly half the articles address, either directly or indirectly, instructional treatments that are helpful to children's learning of multi-digit computation. Eighteen articles discuss research using an observational design and qualitative methods, and eight have a comparative design and statistical methods, but not methods linked to experimental designs. Twenty articles and 18 articles address addition and subtraction, respectively, while seven and four articles address multiplication and division, respectively. Eleven studies used first through third graders, though mostly first through second graders. Seven studies used third through fifth graders. Two studies used only third graders, and six studies spanned multiple grades not described previously. These six studies often spanned four or more grades.

The findings or claims for approaches to instruction on multi-digit computation seem to fit in the spectrum from pre-enactment to enactment. There are findings that range from advocating for what should be emphasized during instruction to the use of conceptual supports and what the teacher should be doing during instruction to allowing and supporting children to invent algorithms. Findings in each of these areas are discussed below.

Several authors claim that instruction should emphasize conceptual understanding and that through building the conceptual

knowledge children build procedural knowledge as well (Brownell & Moser, 1949; Hiebert & Wearne, 1996; Rittle-Johnson & Alibali, 1999). In a comparative study of 1400 third-graders learning subtraction with borrowing, Brownell and Moser found that conceptual knowledge was hard to build upon procedural knowledge, though not impossible. In another longitudinal comparative study of first through fourth graders who were learning multi-digit addition and subtraction, students were divided into two groups: one group of students received alternative instruction emphasizing conceptual understanding via invented algorithms and discussion and the other group received conventional textbook instruction that emphasized direct instruction and practice (Hiebert & Wearne, 1996). These researchers found that conceptual understanding seemed to play a bigger role in the development and adoption of procedures in the alternative instruction classrooms than in the conventional textbook instruction classrooms. They concluded that conceptual understanding makes possible the construction and deployment of meaningful solution procedures and that conceptual understanding must be the goal of instruction.

Research, however, by Rittle-Johnson and Alibali (1999) paints a slightly different picture of the role of conceptual and procedural knowledge in instruction and learning. They concluded that the relation between conceptual and procedural knowledge is not unidirectional, instead, conceptual and procedural knowledge appear to develop

iteratively, with gains in one type of knowledge leading to gains in the other. This study, comparative in nature as well, considered fourth- and fifth-graders addition and multiplication knowledge and the concept of equivalence. Students were assigned to one of three instruction groups, either conceptual instruction, procedural instruction, or control (no instruction).

While there are too few studies to generalize these claims, these three studies do paint a picture of a slightly changing role of conceptual and procedural knowledge in students' learning of concepts and procedures. That is, for children in first through third grades, procedural knowledge evolves out of instruction that emphasizes conceptual knowledge, and for older children, fourth and fifth graders, they benefit from varying the instructional emphases between concepts and procedures.

In several studies by Fuson and colleagues, manipulatives have been found to aid in the development of conceptual understanding and procedural knowledge, and therefore, this research advocates for instruction that meaningfully links concrete quantities to written methods (Fuson, 1990a; Fuson & Briars, 1990; Fuson & Burghardt, 2003; Fuson, et al., 1992; Fuson & Smith, 1997; Fuson, Smith, et al., 1997). In a recent study having a comparative design, 26 high achieving second-graders were assigned to groups of 4-5 children based on pre-test scores (Fuson & Burghardt, 2003). Students worked in groups on

addition and subtraction problems using base-10 blocks. Half the student groups were asked to link block and mark solutions by recording everything they did with the blocks with the digit cards, and the other half of the groups recorded their work on a magic pad. Researchers for this study found that with adult prompts, second-graders were able to translate their concrete procedures into written procedures and that some of these invented written methods were conceptually and procedurally superior to the standard U.S. algorithm. Many incorrect written addition methods were invented because children did not link them to their block solutions. Therefore, pedagogical objects such as base-ten blocks along with teacher monitoring and feedback and the interaction of cooperative-learning groups are essential instructional supports for students' small-group invention of meaningful calculation methods. Furthermore, given a proper social and intellectual environment, children's work in small groups can facilitate their inventions of computational procedures and help them deepen their arithmetic knowledge.

Lampert (1986) also claims that the teacher is important to students' learning. She found while studying her teaching of multiplication to fourth-graders that "[t]he role of the teacher while teaching multi-digit multiplication is to bring students' ideas about how to solve or analyze problems into the public forum of the classroom, to referee arguments about whether those ideas are reasonable, to sanction

students' intuitive use of mathematical principles as legitimate, to teach new information in the form of symbolic structures and emphasize connection between symbols and operations on quantities, and to require students to use their own ways of deciding whether something is mathematically reasonable in doing the work" (Lampert, 1986, p. 339). (Also see Yackel, Cobb, and Wood (1999) for a similar description of the role of teachers in second-graders' learning of addition and subtraction.). Lampert also claims that students need to be asked questions in which answers can be "figured out" not by relying on memorized rules for moving numbers around but by thinking about what the numbers and symbols mean.

Finally, the research advocated for student invented algorithms (Baek, 1998; Hiebert et al., 1996; Kamii & Dominick, 1997; Morrow & Kenney, 1998; Schifter, 1999) and alternative algorithms to the standard ones (W. M. Carroll & Porter, 1998). In fact, Hiebert, Carpenter, et al. (1996) found that instruction in which students see the development of procedures or algorithms as the problem to be solved is a form of instruction in which students will develop high levels of conceptual understanding that are closely connected with their procedural skills.

Weaknesses and gaps. Instructional treatments vary across studies including alternative instruction, direct instruction, and instruction emphasizing concepts, instruction emphasizing procedures, and adult prompts. It is also expected that the quality of instruction

varied as well. Without complete knowledge of the instruction that students received, it is hard to replicate the findings. While this research sheds light on elements of instruction that are helpful to student learning, the uniqueness of instruction makes generalizing the results questionable. Furthermore, the context of instruction varied in other ways: the grade of the students, their ethnicity, achievement levels, and the content focus of instruction. For example, one argument was based on one second-grade class composed of a large Latino population (Hiebert, et al., 1996) — a context not replicated in other studies. In this study, students were engaged in a problem to find the difference in the height of two children — content not replicated in other studies.

In some studies, instruction emphasizing reform ideas was delivered by a specially trained teacher (Hiebert & Wearne, 1996; Kamii & Dominick, 1998) while instruction not emphasizing reform ideas was delivered by the regular classroom teacher (Hiebert & Wearne, 1996). Since instruction varies between classrooms within schools, there is likely more variability in instructional treatments that had been labeled the same by researchers. While instruction delivered by specially trained teachers reduced the variability between classrooms within schools, findings from studies using this mode of instructional treatment are questionable when it comes to replicating the results when classroom teachers are delivering the instruction. Findings are also questionable in studies where instruction was not standard across classrooms since it is

likely that not all students received the same treatment. Furthermore, it is reasonable to question, even with a specially trained teacher delivering the experimental instruction, whether the regular teacher learned about and tried to deliver the experimental instruction. It is not clear how studies controlled for cross over in treatment. More studies are needed that control for variability in instructional treatment and allow for regular classroom teachers to deliver alternative instruction.

Again, research on multiplication and division and on children in the upper elementary grades has been limited in the area. Exactly how instruction should be sequenced to build on children's informal mathematical knowledge is still an open question that requires further study.

Knowledge Teachers Need to Have to Teach Multi-digit Computation

Findings. One article and one book touch on the issue of what teachers need to know to teach multi-digit computation. The authors of the one article address what teachers know. In a correlational study of measures of teachers' knowledge of distinctions between addition and subtraction problem types, children's problem solving strategies, and knowledge of their own students' achievement on number facts and problem solving (Carpenter, Fennema, Peterson, & Carey, 1988). The teachers' ability to predict their students' success in solving different problems was significantly correlated with both measures of students'

achievement. However, their ability to predict the strategies that students would use was not correlated with either achievement measure. The teachers could distinguish some basic differences between the major types of addition and subtraction problems. Most teachers did not appear to have a coherent framework for classifying problems, and they frequently could not articulate the basis for the distinctions they made between problems. Most teachers were familiar with the most frequently used strategies for solving addition and subtraction problems, and they could successfully identify strategies when they observed children using them on videotape. However, they generally did not categorize problems in terms of the strategies that children use to solve them. Many teachers did not seem to recognize the general principle that problems that can be directly modeled are easier to solve than problems that cannot be directly modeled.

Most teachers were reasonably successful in identifying many of the critical distinctions between problems and the primary strategies that children use in solving addition and subtraction problems. However, this knowledge generally was not organized by the teachers into a coherent network that related distinctions between problems, children's solutions, and problem difficulty to one another - relationships that took researchers many years to specify clearly. None of the measures of teachers' general knowledge of problems, problem difficulty, or strategies were significantly correlated with student achievement or even with

teachers' ability to predict either their students' success in solving different problems or the strategies the students use to solve them. Most teachers had the general knowledge needed to predict their students' problems-solving performance and were capable of identifying their own students' strategies when they saw them.

In interviews of U.S. and Chinese teachers discussing their instruction of subtraction with regrouping, Ma (1999) concluded that "three kinds of mathematical knowledge are included in a fully developed and well-organized knowledge package of conceptual understanding: procedural topics, conceptual topics, and basic principles of the subject" (Ma, 1999, p. 23). She found, however, that 77% of the U.S teachers and 14% of the Chinese teachers displayed only procedural knowledge of the topic and that the variation in knowledge of subtraction with regrouping led to different instructional goals.

Weaknesses and gaps. While correlational significance can provide direction, this information does not illuminate what knowledge teacher *need* in order to lead students in learning multi-digit computation. Here, what is needed are carefully controlled studies that rule out possible confounders. The relationship between teachers' ability to predict their students' success in solving different problems and measures of students achievement on numbers facts and problem solving may be related to other measures and in fact it is these measures that are causing the correlational significance.

Additionally, we need to know more about how teachers' knowledge relates to topics within the domain of multi-digit computation and how their knowledge relates to various instructional practices and student achievement. This relationship may of course vary by grade and student composition (SES, achievement levels, parental support). The relationships between teacher knowledge and student characteristics need to be considered as well.

Opportunities for Teachers to Learn to Teach Multi-digit Computation

Findings. This area has been addressed by a collection of articles on teachers engaged in instruction following California's reformation of mathematics instruction and an article resulting from the Cognitively Guided Instruction study. Three second-grade and two fifth-grade California teachers were observed during instruction on multi-digit addition and/or subtraction and multi-digit multiplication and/or division, respectively (Ball, 1990; Cohen, 1990; Cohen & Ball, 1990a, 1990b; Darling-Hammond, 1990; Peterson, 1990a, 1990b; Wiemers, 1990; Wilson, 1990). These California teachers' opportunities to learn the practices called for by the reform policies came in the form of textbooks, district-level materials, and the "Mathematics Framework," not all of which were read by the teachers. Regardless of which reference to the reform policies was used, these teachers learned policies from their current frames' of reference – their pre-existing practice, knowledge, and

beliefs – while working in conditions that lacked time, support, encouragement, and access to new knowledge. Here, teachers needed to learn what was being asked of them in light of their current practice, and unfortunately, opportunities to learn came in the form of written materials, which were filtered through each teacher's frame of reference and understanding of the policies and resulted in no change in practice and no apparent learning.

The research by Fennema, Carpenter, Franke, et al. (1996) reported on teachers' learning to teach in new ways. These researchers worked with 21 first-, second-, and third-grade teachers and their students over a 4 year period. They found that knowledge of children's thinking is a powerful tool that enables teachers to transform this knowledge and use it to change instruction. Here, each teacher came to know more or gained a better understanding of their children's mathematical thinking through exposure to research on children's mathematical thinking and through engaging their students in a variety of problem-solving situations and encouraging them to talk about their mathematical thinking. The results of this learning yielded a change in teachers' beliefs and an increase in student achievement in concepts and problem solving.

Weaknesses and gaps. Not every teacher has the opportunity to work with researchers to learn about current knowledge on students' thinking. Furthermore, does learning about children's thinking and

using it in instruction get harder as the content gets harder? Studies like the one by Fennema and her colleagues need to be replicated with teachers varying in teacher characteristics, school environment, and student characteristics using a more experimental design.

Conclusions

While the teaching and learning of multi-digit computation has been more thoroughly investigated than other mathematical content areas, there are several lines of inquiry that are lacking in knowledge. First, research is deficient on the teaching and learning of multiplication and division. The majority of studies have focused on addition and subtraction and early elementary grades, while far less effort has been put forth investigating the teaching and learning of multiplication and division. Further, additional examination is needed to understand both how children approach multiplication and division and if instructional approaches apply equally to these operations as they do to addition and subtraction. There is a hint of evidence that the interplay between conceptual and procedural knowledge may change as children advance in the mathematical knowledge. Since children can rely on informal mathematical skills for many years and even after formal mathematical instruction occurs, a focus on conceptual knowledge in the early grades may help students move from using informal skills to using formal skills. Then, in the later grades, teachers may need to focus on both conceptual

and procedural skills, iteratively, in order to facilitate learning more advanced mathematics for which informal skills can not be relied upon.

Second, research is needed that includes older elementary children, especially those in fourth- and fifth-grades. Most research has focused on teaching and learning of students in the first three grades. Since teachers reported that work on addition and subtraction continues throughout the elementary grades, there is more to be learned about the teaching and learning of addition and subtraction in later grades. If the lack of inclusion of older elementary students is a subtle way for the field of mathematics education to say that teachers should no longer be concentrating instruction and learning on addition and subtraction in later grades, then research is still needed to understand the reason for this later research and the connection between early investigations and later studies on these topics. As one can expect, unsatisfactory treatment of topics in early grades will lead to continued coverage in later grades. The content of U.S. elementary mathematics has been described as spiral and that its curriculum adds more topics while never dropping early topics. From this review, there appears to be little understanding of why our curriculum is designed in this way. Therefore, greater in depth research is needed on all topic areas and grade levels so that we can move to understanding the whole elementary mathematics curriculum.

Third, there is a lack of connection, coordination, or acknowledgement between the multiple areas of research within the

research on multi-digit computation. For example, research has argued for or found benefit in an instructional use of base-ten blocks and a focus on invented and alternative algorithms. New research should incorporate these findings while testing out new theories. If a new theory, for example, is about teachers learning to teach for understanding, teachers should include base-ten blocks and alternative algorithms in their instruction along with the component that is being studied in the research. Further, there is little argument within the field in general, except of course on the topic of conceptual and procedural knowledge. Research needs to account for theories that have been shown to work by other research and to consider alternative theories. That is, research needs to be cumulative in nature. It needs to build on other research in this field. Research also needs to be coordinated in order to make progress.

Fourth, many of these studies are set in a teaching and/or learning context that does not mirror the teaching and learning context found in the majority of schools in the Unites States. Studies are needed that assess students' performance and instructional practices during class when they are engaged in mathematical activities with their peers and regular teacher. Most studies included in this review used pre- and post-test, interviews, and/or observations, and many times students were assessed, interviewed, or observed individually outside the classroom. Occasionally, students were observed in small groups. These

data collection opportunities do not occur under usual conditions and may affect how students perform and what is said or observed and how the observations are interpreted. Ultimately, the findings may be misleading. For example, Carpenter and his colleagues have been intervening on teachers' practices, an intervention that not all teachers have access to. Fuson and colleagues studied the usefulness of base-ten blocks for students' learning while working in small groups, where students were pulled out of class. If we are to persuade teachers to embark on the hard work of changing their practice, research is needed that shows teachers that the instruction being advocated is possible within the teaching and learning environment and context similar to theirs. Therefore, research is needed that occurs in classrooms as they normally exist, with the interruptions, discipline challenges, and time constraints that teachers manage everyday. This is not to say the current research has little value. Its greatest value is in leading the next phase of research – the "clinical trials" of education.

Finally, most studies are either observational in nature or provide expert opinion favoring a theory, practice or approach. Even the few comparative studies are not truly experimental and are not designed to show a causal effect of an instructional treatment. Further, samples sizes were relatively small and subjects were chosen non-randomly — both of which impact the studies' generalizability. While one study used 1400 third-graders (Brownell & Moser, 1949), this was the exception. More

typical, study designs ranged from one to a half dozen students, classrooms, and/or teachers. In general, the mode of data collection and small sample sizes have not permitted research to move beyond providing expert opinions and the use of small sample statistics. Therefore, research, currently nonexistent, is needed that permits statistical evidence from randomized controlled studies allowing for causal claims in instruction on student achievement.

So, what is known about instruction on multi-digit computation? A lot, but yet not enough. Research on the teaching and learning of multi-digit computation has laid a strong foundation for the next phase of research. At this point in the discussion, it seems more appropriate to ask, what isn't known about instruction on multi-digit computation? In general, there is a need for more research on instruction on multiplication and division, for research to consider instructional approaches that are effective for learning all operations, for study designs to include all elementary grades, for research that takes place in natural educational settings with regular teachers delivering the instruction, for research that builds on prior knowledge, and for study designs that move beyond observational designs and allows for more rigorous analytic methods. Future research should include and allow for the understanding the causal effects of instructional practices on student achievement, how variations in instructional practices arise, and how

teachers' characteristics, particularly their knowledge for teaching, affect their engagement in high quality instructional practices.

Chapter III

Methods

This chapter describes the data and research methods used to compare two instructional approaches used to teach multi-digit computation in the elementary grades. The two approaches differ in their emphasis on multi-digit computation procedures and concepts, two strands of mathematical proficiency. Little is known about the causal effects of these two well-studied instructional approaches in mathematics education on students' learning. This is in part due, on the one hand, to challenges in garnering useful data at-scale on such instructional effects and achievement, and, on the other hand, even when such data is available challenges to measuring instructional approaches and achievement and to drawing valid causal inferences abound. Fortunately, there have been recent methodological advances in causal inferences for non-experimental data which I draw upon in this study and discuss in this chapter. Still, there is little guidance on identifying and assigning labels to classrooms on instructional approaches post enactment. This is new territory. In this chapter, I describe the methods used for gleaning evidence on the annual instructional approach used in study classrooms

from data collected on daily teacher logs. Then separately, I describe the methods used to measure student achievement that get "closer" to the target of instruction than previous studies. In the last section, I describe the methods used to compare instructional approaches and study the causality of these approaches on related achievement measures.

For purposes of clarity, I restate the research questions being addressed in this chapter. The main research question asks, Does the blended instructional approach to teaching multi-digit computation in the elementary grades cause greater learning for all students? To answer this larger questions, I address four questions– (1) do teachers choose their instructional approach based on who they are teaching or is their approach influenced by other factors (i.e. teachers' knowledge, years of experience, professional preparation, gender, and ethnicity; curriculum material; class average ethnic, gender, and socioeconomic composition; and school climate, and (2a) do classes who receive the blended instructional approach learning more than classes receiving the common instructional approach (2b) does the benefit of blended instruction vary by level of prior knowledge, and (2c) are there grade-level differences?

I have organized the chapter into three sections. The first section describes the study from which the data was drawn, the data, and the sample. In the second section, I attend to methods related to measurement and missing data. In the last section, I describe the

analytic methods and models that are used to answer the preceding research questions.

Data

For this dissertation, I use data collected by the Study of Instructional Improvement (SII), originally a large-scale quasi-experimental study of the design, implementation and instructional effectiveness of three of America's widely disseminated CSR programs. SII collected data in two phases from schools and, within each phase, on two cohorts of students.

Phase 1 began in AY 1999-2000 with the entry of 53 elementary schools and Phase 2 added 66 more schools in AY 2000-2001. Within each school, two cohorts of students were studied as they passed from kindergarten through second grade and third through fifth grade. Within each classroom, eight students were selected at random as "target students." If a target student left the school, they were replaced. Teachers of these target students were asked to record information about the instruction they received in mathematics and language arts on the Mathematics Daily Teacher Log and the Language Art Daily Teacher Log, respectively. SII conducted bi-annual — fall and spring — assessments of students' achievement using the CTB McGraw-Hill's TerraNova, mathematics, language arts, and reading sections. Additionally, during each year of the study, SII administered several questionnaires, surveys and interviews to collect data from teachers, school leaders, parents of

target students, and administrators on additional information that may be related to instructional improvement. In this dissertation study, I use data from all sources but with greater emphasis on the Mathematics Teacher Log data and the mathematics items from the TerraNova. In the following section, I describe the data collected using these two instruments. Electronic copies of the instruments are available at www.sii.soe.umich.edu.

Mathematics Teacher Log

SII developed a four page daily mathematics log aimed to measure the mathematics instruction target students received during instruction by their regular classroom teacher. This log was specifically designed for teachers to record mathematics instruction occurring in elementary school classrooms over a given day. On page 1 of the mathematics log, the teacher was asked to record the duration of the instruction or, if applicable, the reason for no instruction. When instruction occurred, teachers were asked to record when eight topics - number concepts; operations; patterns, functions, or algebra; money, telling time, calendar; representing or interpreting data; geometry; measurement; probability; percent, ratio, or proportion; negative numbers; other - were addressed and the extent to which these topics were addressed as being a major focus, minor focus, touched on briefly, or not taught. If teachers recorded a major or minor focus on number concepts; operations; or patterns,

functions, or algebra; they were directed to pages 2 through 4, respectively, of the log to give more details of this instruction. On these pages, teachers are asked to record instruction on number, subtopics, and resources, and on instructional activities and organization. Section A, page 2, focuses on instruction on Number Concepts; section B, page 3, focuses on instruction on Operations; and section C, page 4, focuses on instruction on Patterns, Functions, or Algebra. This study uses items from section B, Operations, which I describe in a later section.

Log data collections were broken into three sessions per logging year—one each in the fall, winter and spring—for about 80 days and about 90% were completed and returned to project staff in usable form. Logging only occurred for the target student of that day and logging was not scheduled for kindergarten.

Of the 89,184 logs collected by SII, 72,852 logs provide records of instruction when school was in session, and both the teacher and student were present at school on the logging day. About 20% of the total logs come from each grade.

Validity of teacher log items. SII researchers developed the mathematics teacher log with attention to accuracy and validity. First, they chose to use a daily teacher log format instead of less frequently administered questionnaires or costly observer records. Daily logs or diaries are considered more accurate since they suffer less from the inaccuracies due to memory distortions. In this case, the daily logs were

completed shortly after the logging events occurred. In all, the use of daily teacher logs, its timing and collection procedure, were all chosen to ensure validity of data collected on instruction practices use in the classrooms under study.

Second, the content of the mathematics log was chosen to represent a range of instructional approaches. Most teachers at the time the log was used were aware of instructional practices recommended in the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989). Researchers "…developed …[log items]… that were agnostic with respect to particular views of good teaching" (Ball & Rowan, 2004, p. 7). Items or groups of items are written so that they do not describe "good" and "bad" teaching or measure teachers' knowledge of reform practices.

Third, these researchers worked through several rounds of instrument pilot testing, incorporating language comprehensible to teachers. In this process they developed a log glossary that provided meaning for terms and examples. For the final log, teachers received training and access to support throughout the study.

Finally, researchers conducted a formal validation study on a pilot version of the language arts log using triangulation methods (Camburn & Barnes, 2004). Depending on the section, they found between 73%-90% comparable agreement between teacher and observer with the more frequently taught content being at the upper end of this range. Observer

to observe agreement was only marginally better at 79%-97% agreement. In the end, the researchers claim that the validation study provided valuable information for revising items, clarifying terms in the glossary, and even eliminating some items.

Mathematics log reliability. SII researchers assessed log item reliability using observer records of 68 mathematics log teachers logging in years three and four of the study. Observations were spread across first through fifth grades, totaling 32, 8, 1, 20, and 7 observations, respectively. Five SII researchers were trained to use the mathematics log. For each observation, two researchers independently read transcriptions of observer records and reported on the mathematics log. The researcher pairs compared their log reports and discussed discrepancies. Once the researcher pair agreed on how to appropriately record the observation, they created one final researcher log called a canonical log. I assessed log item reliability using the canonical and teacher logs. Note that teacher logs were collected as part of the regular study, and no additional or separate logs were collected for this reliability assessment.

Since logs responses are non-continuous or binary, I assess reliability using a Spearman's rank correlation, a non-parametric test of association on the ranked data where -1 and 1 represents perfect agreement and zero no agreement. In Table III.1, I report the correlations for the three focal gateway items and for items in the Operations section,

the focus of this dissertation. For the three gateway items reported, the correlation is moderate, from .30 to .67. Correlations for items in the Operations section range from low, near zero, to moderately high or from -.02 to .89 and for three items, no relationship could be computed because the canonical logs never reported an occurrence while the teacher log reported at least one occurrence.

Many of the log items rarely occur in regular classrooms and the SII researchers expect these items will have low reliability. Because the reliability study included only 68 observations, many of the items that regularly occur in classrooms are likely to have low reliability due to chance, time of year when the observations took place, and grades that were observed. To average out the effects of these factors, I calculated an average correlation across the items in the Operation section, using the absolute value thereby ignoring the direction of the correlation. I found the average correlation on these items to be moderately low at .37 excluding the four items where no correlation was reported.

Table III.1
Spearman rank correlations between teacher and canonical mathematics log items, for focal gateway item and operations section mathematics log items

Log Item (Reference label)	Spearman Rank Correlation
To what extent were the following topics a focus of your work with the target student in mathematics today?	
Number concepts (ml4a)	.50**
Operations (ml4b)	.67**
Patterns, Functions, or Algebra (ml4c)	.30*
Operations (Section B)	
Which operations did you focus on today?	
Addition (mlB0a)	.54**
Subtraction (mlB0b)	.46**
Multiplication (mlB0c)	.34**
Division (mlB0d)	.60**
What were you using in your work on operations?	
Whole Numbers (mlB1a)	.52**
Decimals (mlB1b)	.89**
Fractions (mlB1c)	.21
What did the target student work on today?	
The meanings or properties of an operation (mlB2a)	.18
Basic facts (whole numbers only):	
Methods/strategies for finding answers to basic facts (mlB1b)	.39**
Practicing basic facts for speed or accuracy (mlB2c)	.03
Computation with multi-digit whole numbers, decimals, or fractions:	
Why a conventional computation procedure works (mlB2d)	--
How to carry out the steps of a conventional computation procedure (mlB2e)	.35**
Practicing computation procedures for speed, accuracy, or ease of use (mlB2f)	.34**
Developing transitional, alternative, or non-conventional methods for doing computation (mlB2g)	.40**
Applying basic facts or computation to solve work problems or puzzles (mlB2h)	.35**
Estimating the answer to a computation problem (mlB2i)	.45**

Table III.1 (cont.)

What did you or the target student use to work on the aspects of operations that you checked in Question B2?	
Numbers or symbols (mlB3a)	.55**
Concrete materials (mlB3b)	.41**
Real-life situations or word problems (mlB3c)	.35**
Pictures or diagrams (mlB3d)	.53**
Tables or charts (mlB3e)	.31*
I made explicit links between two or more of these representations (mlB3f)	.48**
What was the target student asked to do during the work on operations?	
Listen to me present the definition for a term or the steps of a procedure (mlB4a)	.23
Perform tasks requiring ideas or methods already introduced to the student (mlB4b)	.54**
Assess a problem and choose a method to use from those already introduced to the student (mlB4c)	.13
Perform tasks requiring ideas or methods not already introduced to the student (mlB4d)	-.05
Explain an answer or a solution method for a particular problem (mlB4e)	.33**
Analyze similarities and differences among representations, solutions, or methods (mlB4f)	--
Prove that a solution is valid or that a method works for all similar cases (mlB4g)	.36**
Did the target student's work on operations today include any of the following?	
Orally answering recall questions (mlB5a)	.31*
Working on textbook, worksheet, or board work exercises for practice or review (mlB5b)	.48**
Working on problem(s) that have multiple answers or solution methods, or involve multiple steps (mlB5c)	-.02
Discussing ideas, problems, solutions, or methods in pairs or small groups (mlB5d)	.25*
Using flashcards, games or computer activities to improve recall or skill (mlB5e)	.30*
Writing extended explanations or mathematical ideas, solutions, or methods (mlB5f)	--
Working on an investigation, problem, or project over an extended period of time (mlB5g)	--

** p <.01, * p <.05

-- Correlation could not be reported due to canonical log never observed.

Responses to the log item measuring time, "How much total time did target student spend on mathematics today? Please include all mathematics instruction the target student received, including routine times such as morning/calendar math, even if [it] took place in another room or by another teacher," might provide reasons for the moderately low reliability, on average. The Pearson correlation between canonical and teachers records on this item is .61.[x] This moderately correlation suggests that teachers might have logged on instruction that was outside the observation time and not included in the observation record.

Despite the moderately low average reliability, SII researchers' use of daily teacher logs, its timing and collection procedure, and their attention to item writing with specific intentions to not sway teachers' responses were all chosen to ensure validity of data collected on instruction practices use in the classrooms under study. Furthermore, their extensive measures to pilot items and logs, revise, and study teachers use of the final log were extensive steps to ensure validity and reliability.

In all, SII researchers' use of daily teacher logs, its timing and collection procedures, and their attention to item writing with specific intentions to not sway teachers' responses were all chosen to ensure validity of data collected on instruction practices use in the classrooms under study. Furthermore, their extensive measures to pilot items and

logs, revise, and study teachers use of the final log were extensive steps to ensure validity and reliability.

Student Achievement Data

I use achievement data collected by SII to construct achievement measures on multi-digit computation. SII used the TerraNova Form A by CTB McGraw-Hill as their achievement assessment. CTB provided to SII students' item specific response data, overall mathematics scores, and scores on several subscales including Number and Number Relations, Computation and Estimation, and Operation Concepts.[xi] The Operation Concepts scale measures both single-digit and multi-digit computation knowledge. Given my interest in studying the effects on achievement on multi-digit computation, I chose not to use the Operation Concepts scale and to create a measure of multi-digit computation from specific items requiring multi-digit computation knowledge. Methods for creating this measure are presented below.

Assessment administrations procedures. SII researchers assessed target students in the fall and spring of each study year. They assigned students to test levels using a rubric developed in consultation with an ETS psychometrican. The goal of the rubric was to minimize the standard error of the scale score. In general, while many levels of the TerraNova are appropriate for assessing students across more than one grade, students were assigned to take the test level associated with their

grade and the time of year the assessment was taken. For example, levels 11 and 12 are the levels best suited for second grade fall and spring assessments, respectively, and levels 12 and 13 for third grade fall and spring assessments, respectively, and so on. If the target student was expected to outperform or under perform on their grade level assessment, an alternate level was recommended. This assessment plan meant that high and low performing students were at risk of taking different test levels from "average" or typical students. This presents concerns for this investigation. Different levels have different items, and, therefore, scores from items on these levels are not comparable. I investigated using methods that might link or make comparable the different levels. I discuss these methods below.

Table III.2 presents the assessment pattern and frequencies by test session, grade, and test form of the collected data. This table only shows the frequencies for second through fifth grades, the grades used in my study. In all, just over 11,000 mathematics assessments were collected during each of Fall and Spring sessions[xii] and about 4000 students were tested in each grade. Since I am interested in measuring student achievement on multi-digit computation and the scale scores provided by CTB McGraw-Hill do not measure multi-digit computation specifically, I use the items requiring multi-digit computation knowledge and students' response data to create scale scores using IRT methodologies. The challenge, however, was whether scores from different test levels could be

comparable. Having comparable scores would render all available data useable, preserving achievement data for potentially high and low achieving students. Alternatively, the loss of scores from high and low performing students means potentially studying the effect of instruction on only "average" students.

I consulted two reports published by CTB McGraw-Hill on the TerraNova that were available at the time of the SII study which provided scant information about the test and items (CTB McGraw-Hill, 1999, October 1997). Advanced technical information on the test and items were unavailable. The Teacher's Guide to the TerraNova claims that "[a]ll [test editions] are tied to a common scale and share a set of linking items." (CTB McGraw-Hill, 1999, p. 4) Furthermore, the reports suggest that test levels overlap in difficulty and mention vertical scaling. Personal conversations with a CTB McGraw-Hill psychometrican confirmed the lack of sufficient published information which would allow items from different forms to be put on the same scale. Furthermore, given the lack of item parameters used by CTB McGraw-Hill for scoring, I pursued several methods of scoring multi-digit items that relied on available information.

First, I pursued methods that preserved cases, such as creating scales related to item difficulty by using items from different levels. However, in the end, this method was insufficient; because of the small

number of multi-digit computation items on each level, there were too few commonalities to accurately define the different levels.

I pursued a second approach that relies on "linking items" or group of "linking students" which are the most reliable methods for linking scores across test levels. Linking items are items that appear exactly the same on more than one test level and have known psychometric properties that facilitate assigning students test scores that are comparable across test levels. The linking items and psychometric properties are not typically available to "end" users, as is the case here.[xiii] Furthermore, after comparing items across Levels 11 through 16, I found no linking items or even similar items that could stand in for linking items. Alternatively, linking students are a group of students who took more than one test level at a single testing session. Here again, this group of students is not available in the SII data or from the available technical information.

The third and final approach I considered relied on scoring students using the item parameters derived from the SII achievement data, the item level data. This approach, however, does not generate comparable scores from different test levels and it also poses concerns that the sample maybe related to the instruction students received. I explore this possibility in a later section after I define the instructional treatment. In the end, this study uses data from test levels taken by the

majority of students, and, therefore, maximizes the sample of useable data.[xiv]

Thus, the focus of my dissertation study is on second through fifth grades using TerraNova Form A Levels 11-15. Table III.3 describes the final analytic sample in terms of test levels. I describe the scales and items in a later section.

Measurement and Missing Data

In this section, I describe how I made use of the "raw" data detailed in the previous section in order to obtain accurate summary measures usable in statistical models. I first discuss the methods used to turn daily teacher log data into annual measures of instructional approach. Second, I describe the IRT methods for scoring the multi-digit computation items. Lastly, in this section, I discuss the methods used to address missing data.

Constructing Measures of Instruction on Multi-digit Computation Procedures and Concepts

This study uses the SII mathematics log items to identify two instructional approaches that place different emphases on procedural and conceptual knowledge and log data to identify classes using the two approaches. I focus on four log items to makes these distinctions. These items are presented in the next section.

Table III.2

Initial Sample: Fall by Spring Crosstabs for students taking the TerraNova by grade and test form (total sample =15,951 students)

		Spring							
TerraNova Form	Missing	Level 10	Level 11	Level 12	Level 13	Level 14	Level 15	Level 16	Total
Fall									
Second Grade									
Missing	1363	0	38	142	17	0	0	0	1560
Level 10	7	1	30	1	0	0	0	0	39
Level 11	165	0	75	1614	0	0	0	0	1854
Level 12	23	0	0	187	7	0	0	0	217
Level 13	1	0	0	0	17	2	0	0	20
Total	1559	1	143	1944	41	2	0	0	3690
Third Grade									
Missing	1184	0	20	64	181	23	0	0	1472
Level 11	1	0	0	0	0	0	0	0	1
Level 12	246	0	0	94	1967	307	0	0	2614
Total	1431	0	20	158	2148	330	0	0	4087

Table III.2 (cont.)

					Spring				
TerraNova Form	Missing	Level 10	Level 11	Level 12	Level 13	Level 14	Level 15	Level 16	Total
Fourth Grade									
Missing	1430	0	1	15	63	153	2	0	1664
Level 10	1	0	0	0	0	0	0	0	1
Level 11	2	0	0	6	0	0	0	0	8
Level 12	24	0	0	0	116	24	0	0	164
Level 13	206	0	0	1	109	1504	23	0	1843
Level 14	26	0	0	0	2	341	25	0	394
Level 15	1	0	0	0	0	1	11	0	13
Total	1690	0	1	22	290	2023	61	0	4087
Fifth Grade									
Missing	1689	0	0	0	21	44	92	16	1862
Level 11	0	0	0	1	0	0	0	0	1
Level 12	0	0	0	0	3	0	0	0	3
Level 13	50	0	0	0	3	191	2	0	246
Level 14	139	0	0	0	0	139	1303	0	1581
Level 15	39	0	0	0	0	0	324	17	380
Level 16	1	0	0	0	0	0	0	13	14
Total	1918	0	0	1	27	374	1721	46	4087

Table III.3
Final analytic sample: Fall by Spring Crosstabs for students taking the TerraNova by grade and test form

	Spring				
TerraNova Form	Level 12	Level 13	Level 14	Level 15	Total
Fall					
Second Grade					
Level 11	932	0	0	0	932
Third Grade					
Level 12	0	914	0	0	914
Fourth Grade					
Level 13	0	0	770	0	770
Fifth Grade					
Level 14	0	0	0	740	740
Total	932	914	770	740	3356

Mathematics log items measuring multi-digit computation. I use items given on page 3 of the Mathematics Teacher Log, called Operations, to identify two instructional approaches to the teaching of multi-digit computation procedures and concepts.[xv] The items in this Operations section ask teachers to report on students' work on computation, including work specifically with multi-digit whole numbers, decimals, and fractions. The first item asks teachers "Which operation(s) did you focus on today?" The teacher can mark addition, subtraction, multiplication and/or division. Instruction involving all of these operations is of interest to this study. The second item asks teachers "What were you using in your work on operations?" Here, the teacher is asked to mark whole numbers, decimals, and fractions. Only lessons

involving whole numbers or whole numbers and decimals are included in this study. Since it is hard to disentangle work on only whole numbers when work on decimals was also marked, I include these lessons in this study. Procedures and concepts used to operate on decimals are similar to those for whole numbers so I include logs when multi-digit whole numbers and decimals are recorded. Since procedures and concepts for operations involving fractions are different from those for whole numbers, logs reporting work involving instruction on fractions is left for future research and not included in this study.

The third item asks "What did the target student work on today?" The teacher is asked to respond to 11 sub-items. Four of these sub-items are the focus of this study. These sub-items are (1) Why a conventional computation procedure works (B2d), (2) How to carry out the steps of a conventional computation procedure (B2e), (3) Practicing computation procedures for speed, accuracy, or ease of use (B2f), and (4) Developing transitional, alternative, or non-conventional methods for doing computation (B2g). These items were designed to measure student work on multi-digit computation procedures and concepts. The glossary statements give some clarity on what each item measures which I restate here:

> **Computation with multi-digit whole numbers, decimals, or fractions:** Computation with multi-digit whole numbers comprises whole number addition, subtraction, multiplication, or division beyond basic facts (e.g., 8 + 14, 23 – 4, 12 x 14, 81 ÷ 3). Computation with decimals or fractions includes any addition,

subtraction, multiplication, or division with any type of decimal or fraction (e.g., .2 + .2 or $\frac{1}{2}+\frac{1}{3}$).

B2d. Why a conventional computation procedure works

Use this category to record work on exploring why a conventional computation procedure works. For instance, when teaching the problem 53 − 29, you or the target student might have used blocks to turn 53 into 40 + 13, explaining that 40 + 13 is another way of representing 53, which makes it possible to subtract the nine in the units column. Or when adding decimals, you might have shown students that lining up the decimal point allows you to combine tenths with tenths, hundredths with hundredths, etc. *If you simply explained the steps or walked through a procedure and did not explain why they work, please mark that in B2e or B2g.*

B2e. How to carry out the steps of a conventional computation Procedure

Use this category to represent work on following steps to complete computation problems. The target student should have worked to master the steps in the procedure, *not yet striving for speed or accuracy.* For instance, you might introduce multiplication with decimals, and the student might work that day to follow the steps of this procedure correctly**.** *If teaching alternative (or non-conventional) procedures, record in B2g.*

Please see the box above for what we mean by conventional procedures.

B2f. Practicing computation procedures for speed, accuracy, or ease of use

Use this category to represent work on helping the target student increase the speed, accuracy, or ease of use in following procedures for computation. The target student might have used flashcards, worksheets, textbooks, games, mental math, or other means of practicing computation procedures.

B2g. Developing transitional, alternative, or non-conventional methods for doing computation

Use this category to represent work on learning, using, or inventing non-conventional methods for computing with whole numbers, decimals or fractions. These include methods that differ from those described above as conventional procedures. Work with nonconventional methods may be informal, such as adding 53 + 19 by “rounding and compensating” (i.e., 53 + 20 = 73; 73 − 1 = 72) or mentally adding the tens before the ones. Other non-conventional methods may help the target student “see” the steps in an operation more clearly, as in these non-conventional methods for multi-digit multiplication and division:

231
x 4
4 (1 x 4)
120 (30 x 4)
800 (200 x 4)
924

17)294
170 (17 x 10)
124
85 (17 x 5)
39
34 (17 x 2)
5

10 + 5 + 2 = 17 **R**5

> With decimal multiplication, the target student might use a non-conventional multiplication procedure and/or might use a non-conventional approach to place the decimal point in the answer. For example, the target student might estimate what would make sense in terms of the size of the answer instead of counting the number of decimal digits.
> With fractions, the target student might do $2\frac{1}{2}$ x 2 by doing 2 x 2 + 2 x $\frac{1}{2}$, for example, or the student might do 3 divided by $\frac{1}{2}$ either mentally or by using pictures and asking how many $\frac{1}{2}$'s go into 3 wholes. (Ball, Cohen, & Rowan, 2002, pp. 11-13)

This concludes the item descriptions for items which were the focus of this study. There are additional items on the teacher log, but these items are not related to instruction on multi-digit computation and not useful to this study.

Connecting log items to emphasis on procedures and concepts. In this section, I document connections between the mathematics log items described above to an instructional emphasis on procedures or concepts. Specifically, I use these items and their definitions to identifying whether the target student worked on multi-digit procedures or concepts. First, two items measure work on procedures - "How to carry out the steps of a conventional computation procedure" (a.k.a. B2e) and "Practicing computation procedures for speed, accuracy, or ease of use" (a.k.a. B2f).

These two items record work focused on "following steps," "master[ing] the steps in the procedure," and "speed, accuracy, or ease of use in following procedures for computation" on conventional procedures. These characteristics embody the kind of work students need to do to gain procedural knowledge. Recall the definition: procedural knowledge refers to a familiarity with mathematical symbols, their appropriate representation and their use *and* to the rules or procedures for solving mathematical problems. Items B2e and B2f exemplify this definition.

Second, the other two items, "why a conventional computation procedure works" (a.k.a. item B2d) and "developing transitional, alternative, or non-conventional methods for doing computation" (a.k.a. item B2g) measure conceptual knowledge. Item B2d is about *why* steps to conventional procedures are sensible or mathematically reasonable. When this item is marked, students are likely to be justifying the steps of conventional procedures and making connections to number concepts. Item B2g is about alternative methods for doing computation. Here, students might also be justifying the steps of alternative procedures or making connections to conventional procedures. Since work where the teacher "*simply explained the steps or walked through a procedure"* is not recorded in these items, marking these items means that students are making connections between steps in procedures and multi-digit number concepts. These items represent instructional emphasis on conceptual knowledge, learning about the *relationships* between computation

methods and making *connections* to the reasoning for the steps in procedures. Recall the definition of conceptual knowledge: conceptual knowledge refers to knowledge rich in relationships and connections between pieces of information. Items B2d and B2g exemplify the meaning of an instructional emphasis on learning concepts.

In all, two log items, B2e and B2f, suggest that an instructional emphasis on procedures occurred, while the other two log items, B2d and B2g, suggest that an instructional emphasis on concepts occurred. Emphasis on either or both procedures and concepts can occur within a single instructional period and vary across instructional periods, and it is this emphasis that I map onto two approaches to instruction on multi-digit computation, previously labeled common and blended instruction. Recall that common instruction places a high instructional emphasis on students' development of procedural skills, with little or no attention to connections with concepts. Classrooms receiving the common instructional approach will log primarily on items B2e and B2f and rarely on items B2d and B2g. Conversely, recall that blended instruction incorporates deliberate opportunities for student work developing conceptual understanding with opportunities for work on procedural skills. For classrooms receiving the blended instructional approach, teachers will regularly log all four items throughout the school year.

Identifying classes using common and blended instruction. In this section, I present the steps I used for identify *which* classes receive

common and blended instructional approaches. I again focus on the four log items measuring daily emphasis on multi-digit computation procedures and concepts and the data collected on these items. Using this information I present my rationale of labeling a class as receiving common or blended instructional approach.

I begin with first studying the rate of daily records for each of the four log items for each grade. Table III.4 restates the four log items and gives the grade average proportion of logs reporting work on each item. Overall, the average proportion increased as grade level increased for all four items. This suggests that with each subsequent grade students are receiving instruction with increasing emphasis on procedures and concepts when multi-digit computation is the focus. Item B2e "how to carry out the steps a of conventional computation procedure" received higher average emphasis compared to the other three items, while item B2g "developing transitional, alternative, or non-conventional methods for doing computation" received the lowest average emphasis. This also means that procedures received more instructional emphasis since B2e measures work on procedures while concepts received less emphasis since B2g measures work on concepts. This concurs with research findings in that typical classrooms emphasize low-level skills and rarely attend explicitly to the important mathematical relationships (National Advisory Committee on Mathematics Education, 1975; Rowan, et al., 2004; Weiss, Pasley, Smith, Banilower, & Heck, 2003).

The rightmost column in Table III.4 restates which—either procedures or concepts — is being measured by the corresponding log item. In the next step I code individual logs as reporting emphasis on procedures or concepts. Therefore, I code a log as including an emphasis on "procedures" if the teacher reported an endorsement of B2e "how to carry out the steps of a conventional computation procedure" or B2f "practicing computation procedures for speed, accuracy, or ease of use" or both. Similarly, I code a log as including an emphasis on "concepts" if the teacher reported an endorsement of B2d "why a conventional computation procedure works" or B2g "developing transitional, alternative, or non-conventional methods for doing computation" or both. From this log coding for procedures and concepts, I estimate for each class the proportion of days that included each emphasis. These proportions provide a measure of how much instructional emphasis a class received on each multi-digit computation procedures and concepts.

In Table III.5, I provide the descriptive statistics for average class emphasis on procedures and concepts by grade.[xvi] The distributions for class emphasis on procedures and class emphasis on concepts are positively skewed except for the distribution of fifth grade class emphasis on procedures. This suggests that there is opportunity for an increase in emphasis on procedures and concepts during instruction on multi-digit computation in most classes. The mean emphasis for procedures varies from .17 for first grade classes to .30 for fifth grade classes, and the

mean emphasis for concepts varies from .12 for first grade classes to .24 for fifth grade classes. These findings are sensible since we know that content gets added to the curriculum and that in many classes earlier content continues to be revisited. For example, Rowan, Harrison, and Hayes (2004) found, "[w]ith respect to redundancy, … that students in third and fourth grades continued to work on addition and subtraction, even as they moved to work on multiplication and division. Moreover, students continued to work on addition and subtraction problems with whole numbers, even as they learned to work with fractions and decimals" (Rowan, et al., 2004, p. 112). The finding that there is variation from grade to grade on emphasis on procedures and concepts, however, suggests that there are elements of instruction that vary between grades but what is not clear is how this variability is related to instructional approach. Still, the low mean class emphasis on both procedures and concepts for first grade compared to second through fifth grades confirms that multi-digit computation is not the focus of their content. This finding also reaffirms my focus on second through fifth grades for this study.

Using these class level proportions, I divide classes into those receiving low emphasis and those receiving high emphasis for both procedures and concepts. I use a cut-point of .2 on the proportion of days endorsing procedures and concepts to divide the classes into high and low emphasis because it is representative of the mean and median

and is generally easy to interpret. For example, I interpret classes above .2 on the measure as having worked, on average when focused on multi-digit computation, at least one day per week on either procedures or concepts, depending on which is being measures. Therefore, I code a class as endorsing a low instructional emphasis on procedures if the class proportion is .2 or less. Conversely, I code a class as endorsing a high instructional emphasis on procedures if the class proportion is greater than .2. I use the same rule for coding low and high instructional emphasis on concepts. The results from this coding show that there are four types of class emphasis as shown in Table III.6, the table of frequencies of classes for each type. There are 407 classes which I indentified as endorsing a low instructional emphasis on procedures and concepts, located in the upper left-hand cell of the table, and there are 459 classes which I indentified as endorsing a high instructional emphasis on procedures and concepts, located in the lower right-hand cell of the table. The remaining cells of the table indentify classes endorsing only high emphasis for either procedures or concepts. For this analysis, I study only the two types of instructional emphasis defined by high emphasis on procedures and low emphasis on concepts (i.e. 268 classes) and high emphasis on both procedures and concepts (i.e. 459 classes), here by referred to as common and blended instructional approaches, respectively.

Table III.7 shows the sample of classes by grade receiving common and blended instructional approaches. I have identified between 169 and 189 classrooms across the different grades and between 31 to 43 percent received a common instructional approach. Therefore, there are enough cases to study instructional effects on achievement and to explore influences on endorsing an instructional approach. This is the sample of classrooms and students in these classrooms that are used to answer my research questions.

Sample descriptives. Table III.8 describes the students and classes by grade in terms of covariates. Student characteristics are fairly similar across the grades. Each grade has between 45 and 50 percent males, 17 and 24 percent Caucasian, 46 and 52 percent African American, and about 20 percent Hispanic. Second grade has slightly higher poverty than the other grades with an average socioeconomic status (SES) of -.03 while the other grades have an average SES between -.10 and -.13. Achievement in reading, language arts, and mathematics increases as grade increases, as expected. For second grade, the average achievement score in mathematics, for example, is 541.87 while the average for fifth grade is 615.54. Proportions of LEP (A.K.A. Limited English Proficiency) and learning disable are similar across grades as well.

Teacher characteristics are similar across grades except for a few noteworthy differences. In terms of teacher's gender, the proportion of male teachers generally increases as grade increases. Second and third

grades have fewer male teachers, 7 and 6 percent, respectively; while fourth and fifth grades have 13 and 22 percent male teachers, respectively. Over 50 percent of teachers are Caucasian with 25 percent African American and about 10 percent each Hispanic and other ethnicity. About two-thirds of teachers in each grade majored in education and about the same proportion have a graduate degree. Between 83 and 87 percent of teachers in each grade hold a permanent or standard teacher certification. On average, teachers in each grade have between 11 and 14 years of teaching experience but have been teaching between six to eight years at their present school. Average course taking and professional development is similar across grades as well. Average scores on content knowledge for teaching, however, are markedly different. Average second and third teachers' scale scores are lower, at .06 and -.01, respectively, while fourth and fifth grader average teacher scores are .27 and .26, respectively.

Table III.8 displays average school characteristics by grade. At the school level, there are few differences across grades. For example, average school enrollment is just under 500 students with 6.45 hours school day. Average schools are comprised of about 50 percent African American students and 46 percent of students come from single parent family. About 10 percent of parents report problem behaviors at home and 6 percent of student are LEP or ESL. Schools are distributed nearly equally across the four whole school reform programs, America's Choice,

Accelerated Schools, Success for All, and comparison schools, studied by SII. On average, 30 percent of schools have an NSF curriculum in use at their school.

Exploring the relationship between test level and instructional approach. My investigation, described in the sections that follow, excludes about 79 percent of the student achievement data collect by the SII study. I exclude these students for several reasons. About 47 percent (n=7490) students were excluded because either or both fall and spring achievement scores are missing. Another 19 percent (n=3032) of students are in classes that did not receive either the common or blended instructional approach and are excluded. The remaining 13 percent are excluded because they took a non-standard sequence of test levels. Since SII researchers assigned students to test levels during their data collection, it is possible that the instructional approach they received may have influenced their assigned test level. For example, fourth grade students receiving the blended instructional approach may have been performing better than peers in common classrooms and assigned to take level 15 in the spring after taking level 13 in the fall. Alternatively, students in common classrooms may retake level 13 in the spring because they were un-performing compared to their peers. This is just one example and instruction may influence assignment to test level in different ways. In this section I explore the relationship between test level and instructional approach and its potential for biasing my results.

I first explore the relationship between Fall test level and instructional approach. Specifically, Fall test level predicts instructional approach. This sample includes all students with a Fall test score. Results from crosstabs and chi-square tests are given in Table III.9. For all grades, the chi-square statistics, given in the rightmost column, are not significant at the .05 level. Therefore, the proportions of student taking a test level are similar across instructional treatments, common and blended, and there is no evidence that test level influences which instructional approach the students received.

In a second analysis, I explore whether instructional approach influences students' Spring test level in two ways. First, I test whether instructional approach predicts Spring test level. This sample includes all Fall test takers and any new students added to the sample in the Spring. Some students in this sample are missing a Spring assessment and their respective test level. Including these students in this exploration allows me to assess whether instructional approach influences their missing Spring assessment or whether poorly performing students were less likely to be assessed. In Table III.10, I give the results from the crosstabs and chi-square tests. Overall, for all grades, the chi-square statistics are not significant at the .05 level. Therefore, the proportions of students receiving common and blended instructional approach are similar across Spring test levels, including missing assessment. For fifth grade only, there is a significant difference in the

proportion of students receiving common and blended instructional approach and taking level 14 only. For the 162 fifth grade students taking level 14, 15.7 percent received the common instructional approach while 10.6 received blended. The difference between these percentages is not exceptionally large but the pattern does different from those for the other test levels. Specifically, at every other test level, the proportion of fifth grade students receiving blended instructional approach was mildly higher than those receiving common which is the reverse pattern to what occurred at level 14. To get a better idea of significance of these findings, I explore the relationship between test taking pattern and instructional approach.

In this second approach I use a measure of the difference between Fall and Spring test levels, calculated as Spring test level minus Fall test level, to assess whether students moved from Fall to Spring without being influenced by their instructional approach. The Spring to Fall test level difference ranges from -1 for down one level to 2 for up two levels. TerraNova test levels can be used to measure student mathematical knowledge at more than one time point, varying both by Fall, Spring, and grade. SII's intention was to assign students to the level that was best suited for individual students. Also, they followed a general protocol of assigning students in the Spring to the test that is one level up from their Fall test level. Therefore, these fourth graders may be taking the next

appropriate level, and this finding is due to chance and not due to their instruction. Therefore, it is not clear whether the fourth grade results in is concerning.

Table III.11 gives the results from crosstabs and chi-square tests for instructional approach predicting change in test level. Given the small number of proportions in question, there is no evidence that test level influences which instructional approach the students received. For all grades the chi-square statistic is not significant. Therefore, overall, the proportions of students in each change category are not significantly different for the two approaches. For fourth grade, there is one category, namely up one level, that has significantly different proportion of students. Of fourth grade students receiving the common instructional approach, about 81.9 percent moved up one test level while about 75.3 percent of those receiving the blended approach moved up one test level. The difference in proportion is not exceptionally large. It does, however, suggest that fourth graders receiving a common approach were more likely to move up one level that their peers receiving the blended approach.

Overall, the results from comparing instructional approach and test level measures suggests there is no clear pattern that instruction influences students' assigned test level. The differences in percentages at fourth and fifth grades are small and biases to results of my investigation are unlikely.

Table III.4

Proportion of logs from the Operations section of the SII Mathematics Teacher Log that focus on multi-digit computation instructional practices (n=72852)

Item (Reference Number)	A Proportion Endorsed either "Touched On" or "A Focus of Instruction"						B Category
	All Grades	Grade					
		First	Second	Third	Fourth	Fifth	
Sample	*72,852*	*14894*	*13984*	*16040*	*15432*	*12502*	
Why a conventional computation procedure works (B2d)	.15	.10	.16	.14	.17	.21	Concepts
How to carry out the steps a of conventional computation procedure (B2e)	.21	.13	.22	.20	.24	.27	Procedures
Practicing computation procedures for speed, accuracy, or ease of use (B2f)	.17	.11	.19	.17	.18	.20	Procedures
Developing transitional, alternative, or non-conventional methods for doing computation (B2g)	.11	.07	.11	.09	.13	.15	Concepts

Table III.5

Descriptive statistics for class proportion of days endorsing procedures (abbr. Procdr) and concepts (abbr. Concpt)

			Grade									
	All Classes (1491)		First (308)		Second (295)		Third (336)		Fourth (305)		Fifth (247)	
	Procdr	Concpt	Procdr	Concpt	Procdr	Concpt	Procdr	Concpt	Procdr	Concpt	Procdr	Concpt
Mean	.25	.19	.17	.12	.27	.20	.26	.17	.27	.21	.30	.24
(SD)	(.19)	(.17)	(.17)	(.15)	(.19)	(.19)	(.18)	(.17)	(.19)	(.16)	(.18)	(.19)
Median	.22	.14	.10	.06	.25	.15	.23	.13	.24	.17	.26	.19
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	1	.94	.87	.83	0.9	0.89	.89	.87	1	.88	.87	.94
Kurtosis	.78	2.08	2.40	4.10	1.24	1.96	.95	2.68	.80	1.69	0.02	1.36
(SE)	(.13)	(.13)	(.28)	(.28)	(.28)	(.28)	(.27)	(.27)	(.28)	(.28)	(.31)	(.31)

Table III.6

Frequency (proportions) of classes assigned to high and low emphasis on procedures and concepts, a crosstabulation (n=1183)

		Procedures		
		Low Emphasis	High Emphasis	Total
Concepts	Low Emphasis	407	268	675
		(.34)	(.23)	(.57)
	High Emphasis	49	459	508
		(.04)	(.39)	(.43)
Total		456	727	1183
		(.39)	(.62)	(1.00)

Table III.7

Frequency (proportions) of classes by grade endorsing high emphasis on procedures and high emphasis on both (n=727 classes)

	Grade				
	Second	Third	Fourth	Fifth	Total
High Emphasis on Procedures (a.k.a. common instructional approach)	80 (.43)	77 (.41)	56 (.31)	55 (.33)	268 (.37)
High Emphasis on Procedure and Concepts (a.k.a. blended instructional approach)	108 (.57)	112 (.59)	125 (.69)	114 (.67)	459 (.63)
Total	188	189	181	169	727

Table III.8
Descriptive statistics on independent variables for students, classes, and schools in analytic sample, prior to multiple imputation

	Grade											
	Second			Third			Fourth			Fifth		
Independent Variables	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Student Level	932			914			770			740		
Classroom Level	180			171			171			158		
Student characteristics												
Proportion male		0.50	.19		.49	.18		.47	.19		.45	.19
Proportion by Ethnicity												
Caucasian		.24	.33		.24	.32		.20	.29		.17	.28
African American		.47	.40		.46	.37		.48	.39		.52	.40
Hispanic		.21	.32		.23	.32		.20	.31		.21	.30
Asian		.04	.15		.04	.12		.08	.24		.05	.17
American Indian		.00	.03		.00	.02		.00	.02		.00	.01
Other ethnicity		.04	.13		.03	.09		.03	.11		.04	.10
SES composite		-.03	.50		-.10	.45		-.13	.45		-.10	.44
Fall TerraNova Scores												
Reading		589.97	25.67		605.05	20.50		625.30	24.00		630.75	21.18
Language Arts		584.02	23.47		597.67	20.68		621.02	21.30		631.60	20.90
Mathematics		541.87	22.48		561.35	21.92		599.53	29.21		615.54	25.67
Percent LEP/ESL (prop.)												
Under 5% of students		.67	-		.70	-		.72	-		.68	-
5%-50% of students		.27	-		.21	-		.20	-		.28	-
Over 50% of students		.07	-		.09	-		.09	-		.04	-
Percent Learning Disabled (prop.)												
Under 5% of students		.61	-		.51	-		.55	-		.57	-
5%-50% of students		.39	-		.49	-		.45	-		.43	-
Teacher Characteristics												

	Grade										
	Second			Third			Fourth			Fifth	
Independent Variables	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Gender (male=1, female=0)		.07	-		.06	-		.13	-		.22	-
Ethnicity (proportion)												
Caucasian		.56	-		.62	-		.51	-		.51	-
African American		.25	-		.23	-		.28	-		.28	-
Hispanic		.09	-		.08	-		.09	-		.07	-
Other		.10			.07	-		.11	-		.14	-
Undergraduate major (education=1, other=0)		.71	-		.67	-		.72	-		.64	-
Graduate degree (yes=1, no=0)		.62	-		.68	-		.67	-		.69	-
Permanent Teacher Certification (prop.)		.84	-		.84	-		.87	-		.83	-
Years teaching		13.35	10.58		14.17	10.64		14.04	10.51		11.92	9.92
Years at present school		7.60	7.20		8.63	8.77		8.02	7.64		6.78	6.45
Number of math courses (prop.)												
No classes		.06	-		.11	-		.11	-		.09	-
1-6 classes		.72	-		.67	-		.69	-		.69	-
7-15 classes		.19	-		.16	-		.18	-		.19	-
16 or more classes		.03	-		.06	-		.02	-		.03	-
Number of math method courses (prop.)												
No classes		.09	-		.16	-		.12	-		.19	-
1-6 classes		.76	-		.72	-		.75	-		.67	-
7-15 classes		.11	-		.09	-		.11	-		.12	-
16 or more classes		.04	-		.02	-		.02	-		.01	-
Math Professional Development (prop.)												
No hours		.08	-		.07	-		.09	-		.13	-
1-10 hours		.69	-		.58	-		.75	-		.67	-
11-15 hours		.13	-		.14	-		.08	-		.20	-

	Grade											
	Second			Third			Fourth			Fifth		
Independent Variables	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
16 or more hours		.10	-		.22	-		.08	-		.10	-
Content Knowledge for Teaching (CKT)		.06	.98		-.01	.93		.27	1.02		.26	.99
School:	86			91			84			82		
Enrollment[a]		495	192		479	188		467	172		477	187
Length of day (in hours)		6.44	.34		6.44	.34		6.44	.37		6.45	.34
Percent African American		51.86	36.37		52.94	36.96		49.57	36.56		51.93	37.74
Proportion single parents		.46	.15		.46	.15		.46	.15		.46	.16
SES composite		-.08	.37		-.11	.34		-.09	.37		-.011	.36
Teacher average CKT		.00	.30		-.01	.28		.00	.29		.00	.30
Woodcock Johnson Mathematics		13.58	1.90		13.44	1.79		13.60	1.93		13.47	1.80
Proportion of grade repeaters		.15	.09		.15	.09		.15	.09		.13	.09
Proportion problem behavior reported by parent		.10	.05		.10	.05		.11	.05		.10	.05
Proportion ESL students		.06	.08		.06	.08		.06	.07		.06	.07
Whole School Reform Involvement (prop.)												
America's Choice		.28	-		.25	-		.24	-		.23	-
Accelerated Schools		.26	-		.24	-		.25	-		.29	-
Success for All		.26	-		.26	-		.25	-		.23	-
Comparison School		.21	-		.24	-		.26	-		.24	-
NSF Curriculum		.29	-		.28	-		.31	-		.27	-

Note. (Prop.) = proportion.

[a] Enrollment estimates are rounded to whole number.

Table III.9

Frequencies (proportions) and chi-square statistics for TerraNova Fall test level predicting instructional approach, by grade

Grade Instructional Approach	TerraNova Test Level Level 10	Level 11	Level 12	Level 13	Level 14	Level 15	Level 16	Total	x^2 Statistic
Second Grade (n=1141)									.73
Common	7 (43.8)	438 (44.3)	51 (40.2)	5 (55.6)				501 (43.9)	
Blended	9 (56.2)	551 (55.7)	76 (59.8)	4 (44.4)				640 (56.1)	
Third Grade (n=1264)									.63
Common		0 (0.0)	605 (48.1)	2 (50.0)				607 (48.0)	
Blended		1 (100)	654 (51.9)	2 (50.0)				657 (52.0)	
Fourth Grade (n=1232)									.22
Common		0 (0.0)	29 (39.7)	307 (33.7)	69 (28.9)	3 (50.0)		408 (33.1)	
Blended		3 (100)	44 (60.3)	604 (66.3)	170 (71.1)	3 (50.0)		824 (66.9)	
Fifth Grade (n=1238)									.20
Common			1 (100)	55 (42.0)	295 (33.3)	76 (35.8)	2 (28.6)	429 (34.7)	
Blended			0 (0.0)	76 (58.0)	592 (66.7)	136 (64.2)	5 (71.4)	809 (65.3)	

Table III.10

Frequencies (proportions) and chi-square statistics for instructional approach predicting TerraNova Spring test level, by grade

Grade	Instructional Approach			χ^2
TerraNova Level	Common	Blended	Total	Statistic
Second Grade (n=1186)				.93
Missing	7 (1.4)	13 (1.9)	20 (1.7)	
Level 11	17 (3.3)	19 (2.8)	36 (2.8)	
Level 12	481 (93.9)	634 (94.1)	1115 (94.0)	
Level 13	6 (1.2)	7 (1.0)	13 (1.1)	
Level 14	1 (0.2)	1 (0.1)	2 (0.2)	
Third Grade (n=1644)				.91
Missing	135 (17.6)	164 (18.7)	299 (18.2)	
Level 11	6 (0.8)	4 (0.5)	10 (0.6)	
Level 12	37 (4.8)	42 (4.8)	79 (4.8)	
Level 13	513 (66.9)	581 (66.2)	1094 (66.5)	
Level 14	76 (9.9)	86 (9.8)	162 (9.9)	

Table I.10 (cont.)

Grade	Instructional Approach			x^2 Statistics
TerraNova Level	Common	Blended	Total	
Fourth Grade (n=1387)				.97
Missing	35 (7.6)	79 (8.5)	114 (8.2)	
Level 11	0 (0.0)	1 (0.1)	1 (0.1)	
Level 12	5 (1.1)	9 (1.0)	14 (1.0)	
Level 13	42 (9.1)	80 (8.6)	122 (8.8)	
Level 14	366 (79.2)	729 (78.8)	1095 (78.9)	
Level 15	14 (3.0)	27 (2.9)	41 (3.0)	
Fifth Grade (n=1308)				.06
Missing	28 (6.2)	60 (7.0)	88 (6.7)	
Level 12	0 (0.0)	1 (0.1)	1 (0.1)	
Level 13	4 (0.9)	14 (1.6)	18 (1.4)	
Level 14	71* (15.7)	91* (10.6)	162 (12.4)	
Level 15	343 (76.1)	671 (78.3)	1014 (77.5)	
Level 16	5 (1.1)	20 (2.3)	25 (1.9)	

* Denotes significant different between column proportions at the .05 level.

Table III.11

Frequencies (proportions) and chi-square statistics for instructional approach predicting change in TerraNova test level, by grade

Grade	Instructional Approach			x^2
Change in Test Level	Common	Blended	Total	Statistic
Second Grade (n=970)				.82
No Change	44 (10.3)	58 (10.7)	102 (10.5)	
Up 1 Level	385 (89.7)	483 (89.3)	868 (89.5)	
Third Grade (n=942)				.51
No Change	13 (2.8)	13 (2.7)	26 (2.8)	
Up 1 Level	411 (89.7)	444 (91.7)	855 (90.8)	
Up 2 Levels	34 (7.4)	27 (5.6)	61 (6.5)	
Fourth Grade (n=886)				.10
Down 1 Level	0 (0.0)	2 (0.3)	2 (0.2)	
No Change	48 (15.8)	118 (20.3)	166 (18.7)	
Up 1 Level	249* (81.9)	438* (75.3)	687 (77.5)	
Up 2 Levels	7 (2.3)	24 (4.1)	31 (3.5)	
Fifth Grade (n=890)				.72
No Change	63 (19.5)	106 18.7)	169 (19.0)	
Up 1 Level	260 (80.5)	460 (81.1)	720 (80.9)	
Up 2 Level	0 (0.0)	1 (0.2)	1 (0.1)	

* Denotes significant different between column proportions at the .05 level.

This concludes the section on measurement related to instructional approach and sample. In the next section, I describe methods used for measuring student achievement on multi-digit computation.

Constructing Achievement Measures on Multi-digit Computation

The main research question I study in this dissertation considers instructional effects on achievement centered on multi-digit computation. In this section, I describe methods used to construct these measures of student achievement. Specifically, I describe the methods used for identifying useable items and for obtaining scale scores for each student.

The first step in creating scores on multi-digit items was to identify useable multi-digit computation items from the CTB McGraw-Hill TerraNova Form A that require either or both procedural and conceptual knowledge of students. Using the mathematics sections of TerraNova Form A Levels 11-15, I selected items from each level that require students to have knowledge of multi-digit computation in order to answer the item correctly.[xvii] I chose items using the following definition of a multi-digit computation item:

> An item is a multi-digit computation item if, in order to obtain the correct answer, it is reasonable to expect that the student must add, subtract, multiply, and/or divide where one or more of the

numerals are either greater than 9 or a number containing non-zero numbers to the right of the decimal.

Using this definition, I identified between 5 and 13 multi-digit computation items per test level, shown in column B of Table III.12. The total number of mathematics items per level ranges from 26 to 47.[xviii] See column C in Table III.12 for this information.

The second step pertains to scoring students' responses on the items selected in the previous step. To create scores or measures of student achievement using the multi-digit computation items, I use a 2 parameter IRT model (Zimowski, Muraki, Mislevy, & Bock, 2003). In column D of Table III.12, I give the IRT model reliabilities which range from about .48 to .81. In general, tests with fewer items yielded a lower reliability while tests with higher number of items yielded a higher reliability. These reliabilities suggest that the 2-parameter model and test items yield scores with moderate to fairly high certainty of accurate scores for students on multi-digit computation.

I use the student achievement scores on multi-digit computation items in causal models described in a later section. In the section that follows, I describe the methods I used for handling missing data.

Table III.12

Frequencies and reliability estimates for multi-digit, by content and test level

TerraNova Level	A Number Concepts	B Operations	C Total	D Reliability Using all data (2 parameter IRT model)
Level 10	NA[a]	0	30	NA[a]
Level 11	5	5	47	
Fall				.603
Spring				.620
Level 12	4	6	26	
Fall				.556
Spring				.601
Level 13	6	6	30	
Fall				.476
Spring				.518
Level 14	5	10	32	
Fall				.581
Spring				.587
Level 15	5	13	32	
Fall				.758
Spring				.701
Level 16	8	11	31	
Fall				.782
Spring				.814

Note. [a] Fairly early on I stopped keeping track of information on Level 10 because this level was not taken by the students used in this study.

Statistical Methods for Missing Data: Multiple Imputation

Using MICE. In the case of missing values, I use MICE (Version 2.3 for R Version 2.10.1) for multiple imputation (Allison, 2000; Gelman & Hill, 2006a, 2006b; Horton & Kleinman, 2007; Rubin, 1996; Wayman, 2003; White, Royston, & Wood, 2011; Yuan). MICE uses multiple imputation by chained equations and generates imputations based on a series of imputation models, one for each variable with missing values. In the first step in the imputation process, all missing values are replaced by sequential regressions with values from simple random sampling with replacement from the observed values. For the next step, the first variable with missing values, say $x1$, is regressed on all other variables $x2, \ldots, xk$, restricted to individuals with the observed $x1$. Missing values in $x1$ are replaced by simulated draws from the corresponding posterior predictive distribution of $x1$. The process continues with the next variable with missing values, say $x2$, which is regressed on all other variables $x1, x3, \ldots, xk$, restricted to individuals with the observed $x2$, and using the imputed values of $x1$. Missing values in $x2$ are replaced by draws from the posterior predictive distribution of $x2$. The process is repeated for all other variables with missing values and the first "cycle" ends. In order to stabilize the results, the procedure is usually repeated for many cycles to produce a single imputed data set, and the whole procedure is repeated m times to give m imputed data sets.

Multiple imputation on SII data. For multiple imputation purposes, I intended to include many more variables in the imputation procedure than those variables planned for use in the propensity models. However, I found that issues of multicollinearity were abundant and persistent among potential variables. I was able to eliminate some instances of the multicollinearity using principle component analysis where by including only the first principle component for a related group of variables. Generally, the first principle component describes the greatest amount of variation in the group of variables as compared to second and third components, etc. For example, since variables measuring student achievement on mathematics, language arts, and reading were significantly correlated, I was able to combine and describe their common variable using principle component analysis. I use the results from the first principle component in place of the three achievement variables in the propensity score model. To further reduce issues of multicollinearity, I also combined independent variables when the group or first principle component was interpretable. Still, I found that multicollinearity persisted, mainly due to the missing pattern in the data. For this reason, in the end, I only included those variables considered for the propensity score models in the multiple imputation models. With this smaller subset of available covariates, I was able to run the imputation model. I leave improvements to imputation techniques for future research.

In Table III.8, I list the variables used for imputation. Five data sets were imputed. This was the default setting in MICE. From consulting imputation experts, I learned that imputing between five and ten data sets is recommended. The idea is that fewer than five data sets are insufficient for representing the uncertainty of the imputed data and more than ten data sets results in redundant information. Since there was no information suggesting that five data sets were insufficient for this study, I used the default of five imputed data sets.

Since MICE 2.3 does not handle multilevel data, school and classroom level variables were imputed by separate imputation models. All classroom level variables, continuous and non-continuous, were centered at their respective school means (Schafer, 1997; Shin & Raudenbush, 2007, 2010). To do so, for each classroom level variable I subtracted the average school specific response. I ran the imputation on the centered classroom covariates. Following imputation, I then un-centered the variables by adding the average school specific response for each variable. When all classroom level responses within a specific school were missing, the school average was also missing. In this case, I used the grand mean or the mean across all available responses. No adjustments were made to the school level variables, but for imputation I included the imputed classroom level variables. Once I obtained the five imputed data sets for classroom and school level covariates, I checked all covariates to ensure that the imputed values were reasonable by

comparing the mean, standard deviation, minimum, and maximum of the imputed data with the non-imputed data. The imputed data and the non-imputed data should have similar or near equal descriptive statistics if imputation was sufficient. Following imputation, I proceeded with the propensity models discussed in the next section.

MICE assumes that missing data is missing at random (MAR), the probability of data being missing does not depend on the unobserved data, conditional on the observed data (Little & Rubin, 2002). Other forms of missingness include missing completely at random (MCAR) where the probability of data being missing does not depend on the observed or unobserved data and missing not at random (MNAR) where the probability of data being missing does depend on the unobserved data, conditional on the observed data. Departures from MAR can arise when larger fractions of data are missing. One way to deal with this is to include many variables in the imputation model in an effort to make MAR more plausible.

Descriptive statistics for the imputed data are given in Appendix B. Since there are no large differences between descriptive statistics, I concluded that the imputation was successful.

Analytic Models

I begin this section by first presenting the methods used to study *who* uses or receives the blended and common approaches and

conditions affecting their use. These are a set of predictive models including a set of covariates focused on teachers' knowledge, years of experience, professional preparation, gender, and ethnicity; curriculum materials; class average ethnic, gender, and socioeconomic composition; and school climate. In the next section, I expand the predictive model and develop the propensity score models. These models include far more covariates than the previous models. Finally, the last set of models address my research questions about the causal effects of instruction on achievement situated in multi-digit computation. Recall that this dissertation assesses the casual effect of the blended instructional approach compared with the common approach on student achievement measuring multi-digit computation knowledge using the teacher log and achievement data previously described. I hypothesize that students should learn more and have greater success in classrooms receiving blended instruction than if had they been in classroom receiving common instruction.

Predictive Models

What affects the enactment of a common or blended instructional approach while teaching multi-digit computation? In this section, I present the methods and measures used to answer this research question. In particular, I select several key independent variables measuring teacher, class, student, and school characteristics for

inclusion based on current knowledge of instructional. Many agree that instruction is the interaction among teachers, students, mathematics, in contexts (Cohen & Ball, 1999; National Research Council, 2001). Furthermore, summaries of research suggests that "…quality of instruction is a function of teachers' knowledge and use of mathematical content, teachers' attention to and handling of students, and students' engagement in and use of mathematical tacks" (National Research Council, 2001, p. 315). It is this research that guides my selection of variables to be included in the predictive models.

Specifically, I consider 16 variables measuring teacher, student, class, and context characteristics and 10 variables measuring school characteristics for inclusion in the final model which were selected for their theoretical merit in influencing the use of instructional approach. In the next several paragraphs I describe my reasoning for their inclusion. I use variable selection procedures — assess bivariate relationships and enter variables using stepwise, forward, and backward selection — to develop parsimonious models.

First, in order to assess the relationship between teachers' knowledge and years of experience on instructional approach, I consider a measure of teacher's content knowledge for teaching (KNOW) drawn from items on the teacher questionnaire. Each of the four years of the SII study, teachers were asked to respond to a small number of questions about their content knowledge related to mathematics teaching in the

elementary grades. Once the SII study was complete, SII researchers combined the item responses from across the four years to create a single measure of teachers' content knowledge for teaching. I use this measure to represent teachers' knowledge and how well prepared they are to work with students on the content. As shown in Table III.8, fourth and fifth grade teachers have higher content knowledge for teaching than second and third grade teachers with means at .27 and .26, respectively, compared to .06 and -.01, respectively. The standard deviation on this measure is about the same across the grades with third grades being a little less (SD = .93). Since, for many teachers, learning to teach and gaining skills to work with students on the content takes place from on-the-job experience, I also include the first principal component from two items – number of years at current school and number of years teaching – on the teacher questionnaire measuring years of experience (YEARS). On average, depending on the grade, teachers have about 1 to 14 years of experience and spent between 6 to 9 years at their present school.

I consider four variables representing teachers' professional preparation and professional preparation for using the mathematics content for inclusion (Lubienski, 2002). From items on the annual teacher questionnaire, I consider binary measures of whether teachers' undergraduate field was in education (UNDERED), whether they obtained a graduate degree (GRADED), and whether they hold permanent or standard certification (CERT). I also consider controlling

for the number of mathematics professional development hours in mathematics and language arts the teacher completed (PD) and the number of courses taken in mathematics and mathematics methods. See Table III.8 for response categories and descriptive statistics. Again, depending on the grade, between 64 to 72 percent of teachers have an undergraduate degree in education, between 62 and 69 percent have graduate degrees, and at least 83 percent hold permanent or standard teaching certification. More teachers marked having completed between 1-10 professional development hours and 1-6 mathematics and mathematics methods courses. I summarize the information from variables measuring course taking using principal component analysis. The final course variable (COURSE) explained 69 percent of the variance.

Using another group of covariates, I explore the effects of class context through variables measuring teacher and students' gender and ethnicity. I consider including male teacher (TMALE), Caucasian teacher (TCAUC), class average proportion of Caucasian students (SCAUC), class average proportion of African American students (SAA), class average proportion of Hispanic students (SHISP), class average proportion of students of other ethnicity (SOETHN), and class average proportion of male students (SMALE). Overall, fourth and fifth grades have more male teachers, 13 and 22 percent, respectively, than second and third grades, 7 and 6 percent, respectively. About half the students are male, but the proportion decreases as grade increases. More than half the teachers are

Caucasian while between 17 and 24 percent of students are Caucasian, depending on the grade.

Since research has found evidence that teachers make instructional choices based in part on student income status (Lubienski, 2002; Page, 1991; Rist, 1970a, 1970b; Rosenthal & Jacobson, 1968), I also consider including a measure of students' socioeconomic level with the context variables. Using data from the parent questionnaire, SII researchers created a composite SES measure from data collected on mother's highest level of education, mother's occupational status, father's highest level of education, father's occupational status, and household income level. Bivariate correlations of SES with measures of families without enough clothing, emphasizes counting money, and practices percent and multiplication with child, and a measure of number of books in the home are moderately large. I use the SES composite as a measure of students' SES and consider it more generally as a measure of family resources.

SII collected student achievement data on mathematics, reading and language arts. In all grades, student achievement was generally highest in reading, ranging from about 589 for an average second grader to about 631 for an average fifth grader. Average language arts achievement ranged from about 584 to 632 for second through fifth grades, respectively. Average class achievement was lowest in mathematics, ranging from about 542 to 616 in these same grades. From

this data, I created class average achievement measures. Given that these achievement variables are highly correlated, I summarized this formation into one achievement variable (ACHIEVE) using principal component analysis which explains 87 percent of the variance.

Lastly, I consider including, at the classroom level, two variables that measure context in terms of the proportion of students being non-standard learners — percent LEP or ESL students (LEP) and percent of learning disabled students (LD). In most classes, less than five percent of students are considered to be in either of these categories.

In my predictive models, I consider controlling for school context associated with mathematics materials in use. The first of the two independent variables is associated with school reform programs. SII chose schools for their engagement in Whole School Reform, particularly three of the more widely used programs —America's Choice (AC), Accelerated Schools (AS), and Success for All (SFA). About a fourth of the original 115 schools were engaged in each of these programs, and the remaining fourth were chosen as comparison schools (COMP). Much of the guidance and materials provided by the reform programs were focused on reading instruction, however, AC and SFA included mathematics materials. To control for and test the effect of the context and mathematics guidance provided by AC and SFA only, I construct a binary variable (ACSFA) indicating school involvement with either of these two programs. In a second variable, I control for mathematics

curriculum materials in use at the school. Since curriculum materials may or may not provide support for or influence instructional choices, I consider a summary variable of curriculum available in all models. From phone interviews with school personal, SII researchers collected data on mathematics curriculum in use. From this data, I created a binary variable on whether an innovative mathematics curriculum supported by the National Science Foundation (NSF) was in use during any of the four years. For this variable, NSF, a 1 represents that either *Everyday Mathematics* (University of Chicago School Mathematics Project, 1999), *Investigations in Number, Data, and Space* (Cory, 1995), or *Math Trailblazers* (Becker & Morgenthaler, 1998) were available at the time of the study.[xix] In about 30 percent of schools, an NSF curriculum was in use at some point in the four year study.

At the school level, I also consider including two variables that describe the school context in which instruction and learning takes place. Collected from the School Characteristics Inventory, one variable is a measure of the size of the school, or number of students enrolled at the school (ENROLL) and a second variable measures the length of the school day (LENGTH). On average, schools in this study have just under 500 students and in session for nearly 6.5 hours per day.

Another six variables describe student and family context or "neighborhood" in which the school resides. Many of these variables are aggregates of variables included at the classroom level. These variables

include percent African American students (AAS), proportion from single parent families (SINGLE), average family socioeconomic status (AVESES), proportion of students with problem behaviors (BEHAVE), and average kindergarten achievement measure (WOOD). Finally, I consider one final variable, one that measures school average content knowledge for teaching, which may measure teachers' colloquial support for teaching.

Since, during model development, including fewer covariates leads to more precise parameter estimates, I eliminate variables based on statistical evidence. Bivariate relationships of these covariates with treatment are given in Appendix B. Significant bivariate relationships are suggestive of inclusion while collinearity suggests that a variable(s) can be excluded. Note that if two covariates are highly correlated, then there is no need to include them both in the final model. Large standard errors can also be evidence of collinearity. In order to make grade level comparisons, I include the same variables in all models. Since I found that, for this data, coefficients and standard errors were fairly consistent for models having *n* covariates to models having *n-1* covariates, I eliminated covariates that were non-significant in any grade level model. For variables included at the school level, level 2, I also include the corresponding classroom level, level 1, variable even if it was not significant in any of the grade level models.

The final models include 7 classroom-level covariates and 6 school-level covariates which are defined in the model specifications below. At

the classroom level, I removed variables measuring teachers' professional preparation, proportion of non-standard learners, and class average SES from the model due to lack of statistical support, while at the school level, I removed a measure of student behavior due to large standard error and likely collinearity. I also left out variables measuring enrollment, proportion of singles families, and average SES due to lack of statistical support.

For the final predictive model, I use a two-level logistic regression model similar to the one presented above for estimating the propensity scores. Again, the subscript *j* denotes the level-1 units, classrooms, and *k* denotes the level-2 units, schools, such that, I let Y_{jk}, an indicator variable, take on a value of 1 if classroom *j* from school *k* endorsed a blended instructional approach and 0 if classroom *j* from school *k* endorsed a common instructional approach. As before, μ_{jk} denotes the probability that $Y_{jk} = 1$, which varies randomly across classrooms. Therefore, when conditioning on this probability, $Y_{jk} \mid \mu_{jk}$ ~ Bernoulli[xx] and a logistic regression model. The model at level 1 accounts for predictable variation within classrooms across schools. It views the log-odds of finding an emphasis on a blended instructional approach for classroom *k* as depending on aggregate student, class, and teacher characteristics. All variables were entered grand-mean centered which

aides the interpretation of the coefficients. Therefore, the level 1 equation is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = \beta_{0k} + \beta_{1k}(SMALE)_{jk} + \beta_{2k}(SAA)_{jk} + \beta_{3k}(SHISP)_{jk} + \beta_{4k}(ACHIEVE)_{jk} + \beta_{5k}(TMALE)_{jk} + \beta_{6k}(TCAUC)_{jk} + \beta_{7k}(KNOW)_{jk} \quad (1)$$

where

β_{0k} is the average adjusted log-odds for classes emphasizing a blended instructional approach for school *k*;

β_{1k} = the average adjusted log-odds for classes emphasizing a blended instructional approach associated with a class average proportion of male students (SMALE) within school *k*;

β_{2k} = the average adjusted log-odds for classes emphasizing a blended instructional approach associated with a class average proportion of African American students (SAA) within school *k*;

β_{3k} = the average adjusted log-odds for classes emphasizing a blended instructional approach associated with a class average proportion of Hispanic students (SHISP) within school *k*;

β_{4k} = the average adjusted log-odds for classes emphasizing a blended instructional approach associated with a class average achievement (ACHIEVE) within school *k*;

β_{5k} = the average adjusted log-odds for classes emphasizing a blended instruction approach associated with a class having a males teacher (TMALE) within school *k*;

β_{6k} = the average adjusted log-odds for classes emphasizing a blended instructional approach associated with a class having a Caucasian teachers (TCAUC) within school *k*; and

β_{7k} = the average adjusted log-odds of classes emphasizing a blended instructional approach associated with a class teachers' mean response for teacher knowledge (KNOW) within school *k*.

The level 2 model accounts for variation in the log-odds between schools on blended instructional approach. Therefore, the level 2 equation is

$$\pi_{0k} = \gamma_{00} + \gamma_{01}(LENGTH)_k + \gamma_{02}(AAS)_k + \gamma_{03}(AVEKNOW)_k + \gamma_{04}(WOOD)_k + \gamma_{05}(ACSFA)_k + \gamma_{06}(NSF)_k + u_{0k}, \quad (2)$$

where

γ_{00} is the average adjusted log-odds on emphasizing a blended instructional approach across schools;

γ_{01} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with school *k* having average length of school day (LENGTH);

γ_{02} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with an average class proportion of African American students (AAS) in school *k*;

γ_{03} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with school *k* average teacher content knowledge for teaching mathematics (AVEKNOW);

γ_{04} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with average kindergarten students' Woodcock Johnson mathematics score (WOOD) in school *k*;

γ_{05} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with school *k* using an AC or SFA mathematics materials (ACSFA);

γ_{06} = the average adjustment in log-odds on emphasizing a blended instructional approach associated with using an NSF mathematics curriculum (NSF) in school *k*; and

u_{0j} = the random error associated with school *k* after adjusting for class and school level covariates.

This concludes the section on predictive models. In the next section, I extend these models for use in the propensity score methods.

Propensity Score Methods and Models Supporting Causal Inferences

The main research question of this dissertation addresses the causal effect of the common instructional approach versus blended

instructional approach or treatment. Throughout this discussion, I characterize the treatment using a binary indicator if $Z_{jk} = 1$ if the jth classroom in the kth school received the blended instructional approach and $Z_{jk} = 0$ if the jth classroom in the kth school received the common instructional approach. Had this study actually taken place and not used data drawn from non-experimental study, classrooms would have been assigned to a treatment using an experimental design. That is, each classroom would be randomly assigned to receive a common or blended approach to instruction, control or treatment, respectively. The average gain in achievement on multi-digit computation, then, would be computed for the blended, or treatment, and common, or control, classrooms, denoted $E[Y|(Z = 1)]$ and $E[Y|(Z = 0)]$, respectively, and the difference between the two averages, denoted $E[Y|(Z = 1)] - E[Y|(Z = 0) = E[Y(1)|(Z = 1)] - E[Y(0)|(Z = 0)]$, would be an unbiased estimate of the average causal effect of the blended approach. This estimate is unbiased because random assignment ensures that confounding effects from variables such as characteristics of teachers, students and classrooms are unrelated to treatment assignment or receiving the blended instruction and outcome or achievement on multi-digit computation. In the absence of random assignment, such a difference between the two group averages cannot be regarded as an unbiased estimated of the causal effect without some additional statistical controls, as is the case in this study. For this reason, I approach this study using causal

inference techniques that will yield scientifically-based answers. I use Rubin's Causal Model (Holland, 1986; Rosenbaum & Rubin, 1983; Rubin, 1978) for the causal analysis and propensity score stratification in the statistical models. In the next few paragraphs, I will explain how the propensity scores are used in the causal analysis and then present the propensity score models. Following this section, I will present the causal models.

Rubin's causal model. Each class, j,*k,* whether a case or control, has two potential outcomes $Y_{jk}(Z = 1)$ (if case) and $Y_{jk}(Z = 0)$ (if control). The causal effect of the treatment is the difference between $Y_{jk}(Z = 1)$ and $Y_{jk}(Z = 0)$ (or $Y_{jk}(Z = 1) - Y_{jk}(Z = 0)$) for each class. Since the class either belongs to the control group or the case group, it is impossible to observe both $Y_{jk}(Z = 1)$ and $Y_{jk}(Z = 0)$ for each class under the same conditions, the "Fundamental Problem of Causal Inference." Furthermore, if $Z_{jk} = 1$ is the treatment applied to the *jkth class,* $Y_{jk}(Z = 0)$ is the counterfactual outcome and cannot be observed; likewise, if $Z_{jk} = 0$ for *jkth class,* $Y_{jk}(Z = 1)$ is the counterfactual outcome and unobserved. However, we can estimate the average causal effect of a treatment in a population under the assumption that the potential outcomes of treatment are conditionally independent of treatment assignment given the covariates.

This problem of causal inference is directly applicable to the study of causal effects of instruction. If class *jk* received a blended instructional

approach, with $Z_{jk} = 1$ denoting the treatment approach was applied to this class, we would observe the classes outcome in terms of achievement at the end of the school year, represented by $Y_{jk}(Z = 1)$. Given the potential of an alternative treatment, the common instructional approach, causal questions loom. For example, what if this class had received the common instruction approach? If this class would have exhibited a better learning outcome as a results of the common instructional approach, represented by $Y_{jk}(Z = 0)$, $Y_{jk}(Z = 1) - Y_{jk}(Z = 0)$ would be negative, suggesting a detrimental effect of blended instructional approach on this class's academic growth. However, once the class receives the blended instructional approach during an academic year, the alternative treatment or common instructional approach, is no longer available to them. Therefore, the outcome $Y_{jk}(Z = 0)$ is the counterfactual outcome and can never be observed, and questions related to instruction approach remain unanswered. This is the main reason why policies on instructional approach are absent and instructional guidance remains vague.

While it is impossible to estimate $Y_{jk}(Z = 1) - Y_{jk}(Z = 0)$, we can estimate the population average causal effect under some assumptions, denoted as $E[Y(Z = 1)] - E[Y(Z = 0)]$, where $E[.]$ denotes the expected value for the population. This expected causal effect can be interpreted as the difference between the population average potential outcome if blended instructional approach ($Z = 1$) is applied to the whole population

of classes and the population average potential outcome if the common instructional approach ($Z = 0$) is applied to the whole population of classes. Answers to related population questions are of particular interest to educators and educational researchers who advocate that all students benefit from a blended instructional approach while others advocate that effect varies across subgroups. This causal question, therefore, has direct policy implications on whether blended instructional approach offers any benefit at all, on average, to classes or whether the benefit varies across sub-populations of classes.

Given the need to explicate the expected causal effect of blended instructional approach, I now turn to discussing the supportive statistical techniques.

Propensity score methods. In observational studies, assignment to treatment is not randomized. Therefore, potential outcomes, $Y(Z = 1)$ and $Y(Z = 0)$, are no longer independent of treatment assignment and the researcher must be concerned about pre-treatment characteristics of treated and control groups (e.g. teacher and classrooms characteristics) that might be related to the two potential outcomes, $Y(Z = 1)$ and $Y(Z = 0)$. The researcher must infer the treatment assignment from the available data, assuming that all the relevant information has been collected. In most observational studies some of the pre-treatment characteristics get measured while others do not. The characteristics of most concern are those pre-treatment characteristics that are predictive

of potential outcomes $Y(Z = 1)$ and $Y(Z = 0)$ and the treatment assignment $Z = 1$ and $Z = 0$, also called confounding variables. Using the "strong ignorability assumption" (Rosenbaum & Rubin, 1983), the research can ignore non-confounding variables as long as he controls for confounding variables, denoted by X. Therefore, the conditional probability of assigning each class, as is the case in this study, to a particular instructional approach can be expressed, as a function of X, as $e(X) = Prob\,(Z = 1|X)$, where $e(X)$ is called the propensity score. This is read as the propensity score equals the probability of receiving the blended instructional approach given or conditional on the covariates. By including the propensity score which summarizes the information that the covariates carry about treatment assignment, the regression model will yield valid causal inferences.

Estimating the propensity score. When there is only one or only a few confounding covariates, these can be included in the final causal model. However, more often, there are many confounding variables and problems with multicollinearity arise. In such cases, the propensity score model given above which estimates the propensity score is useful. This method replaces a large number of pre-treatment variables with a summary score, the propensity score which contains all the necessary information about treatment assignment from the covariates. Such scores can then be used in causal models to yield valid inferences. This study considers only two groups where the treatment group received the

blended instructional approach and the control group received the common instructional approach. Generally, in the case of a binary treatment, the propensity score is obtained by predicting treatment group membership from the pre-treatment covariates using a logistic regression model. I extend this general case of the propensity score model to a multilevel propensity score model. I use two-level logistic regression models with classrooms nested within schools. Teacher and classroom pre-treatment covariates are included at the classroom level and average teacher and classroom and school pretreatment covariates are included at the school level, see Table III.13 for the specific covariates considered for this analysis.

Selection of model covariates. Model covariates were selected from the variables included in the multiple imputation models discussed in the previous section. In order to not over specify the propensity score models and still have balance between the treatment groups on the observed pre-treatment covariates, I took steps to select the smallest subset of these variables. For each grade, I determined the variables to be used in the propensity score models by testing for significant bivariate relationships using t-tests and through variable selection techniques such as stepwise, forward, and backward regression. Due to imputation, I have five classroom and school level data sets for each grade and my goal here was to identify the subset of observed pre-treatment covariates that hopefully balance on all pre-treatment covariates between treatment

groups for all five imputation data sets. To achieve this, I first identify the appropriate subset of variables for the propensity score model on one data set for each grade and test all possible pretreatment covariates using t-tests and all conditional associations on each pre-treatment covariate between treatment groups, controlling for strata. I then repeat these tests of balance for the other four sets of data. For all grades, I was able to identify a small number of observed pretreatment covariates which I use in the propensity models and successfully balance the two groups on all covariates across the five imputes data sets. In Table III.13, the four right columns, I give the pre-treatment covariates used in the grade-level propensity scores models.

Table III.13

Pre-treatment covariates available for propensity score models, source of measurement, variables selected for PS models by grade

Variable	Source	Grade 2nd	3rd	4th	5th
Classroom Level					
Student Characteristics					
Proportion Males	Parent	X		X	X
Proportion by Ethnicity	Questionnaire				
Caucasian					
African American					
Hispanic					X
Asian			X		X
American Indian				X	
Other ethnicity		X			
Socioeconomic status composite					
First principal component of average achievement	TerraNova				X
Reading					
Language Arts class					
Mathematics[a]					
Percent of LEP or ESL students	Teacher				
Percent of learning disabled students	Questionnaire				
Teacher Characteristics					
Male teacher	Teacher			X	
Ethnicity	Questionnaire				
White				X	
African American					
Undergraduate major field of study in education	Teacher				X
Graduate degree	Questionnaire				

Variable	Source	Grade 2nd	3rd	4th	5th
Permanent or standard certification					X
First principal component of years of experience					
Career experience in years					
Years at current school[b]					
First principal component of 4 items on university content and methods courses		X			
First principal component of math, LA PD hours					
Measure of Content Knowledge for Teaching					X
School Level					
Enrollment	School				X
Length of day	Characteristics Inventory	X	X		X
Percent African American students	Parent		X		X
Covariates on families attending school	Questionnaire				
Proportion of single parents					
Socioeconomic Status composite					
Proportion of families without needed clothing					X
Child practices counting money at home					X
Child is often read to at home				X	
Number of books in home					X
Proportion of problem behaviors reported				X	X
Proportion of grade repeaters					X
Teacher average Content Knowledge for Teaching	Teacher				
Proportion of students in ESL	Questionnaire	X			
Average Woodcock Johnson Mathematics	Woodcock Johnson		X		

Variable	Source	Grade 2nd	3rd	4th	5th
Whole School Reform Involvement	School Characteristics Inventory				
American's Choice					
Accelerated Schools					
Success for All			X		
Comparison School					
NSF Curriculum in use	Phone consultation with school personnel	X			

[a] For principal component of mathematics, reading, and Language arts achievement, .87 variance was explained.
[b] For principal component of years of experience, .85 variance was explained.
[c] For principal component of number of content and method courses, .69 variance was explained.
[d] For principal component of mathematics and Language Arts professional development, .82 variance was explained.
[e] Empirical Bayes estimate was used.

The propensity score model. Next, I describe the logistic regression propensity score models generally, since the models vary by grade. In order to be consistent with the causal models and predictive models described later in this chapter, I use the subscript *j* to denote the level-1 units, classrooms, and *k* to denote the level-2 units, schools. Specifically, let Y_{jk} be an indicator taking on a value of 1 if teacher j from school *k* endorsed a blended instructional approach. Thus, $Y_{jk} = 0$ if teacher *j* from school *k* endorsed a common instructional approach. Let μ_{jk} denote the probability that $Y_{jk} = 1$, and this probability varies randomly across teachers. Therefore, when conditioning on this probability, $Y_{jk} \mid \mu_{jk}$ ~ Bernoulli[xxi] and the appropriate model is a logistic regression. The model at level 1 accounts for predictable variation within teachers across schools. It views the log-odds of finding an emphasis on a blended instructional approach for teacher *k* as depending on aggregate student, class, and teacher pre-treatment characteristics. Therefore, the level 1 equation is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = \beta_{0k} + \beta_{1k}(COV1)_{jk} + \beta_{2k}(COV2)_{jk} + ... + \beta_{19k}(COV19)_{jk}, \quad (3)$$

where

β_{0k} is the class average adjusted log-odds of emphasizing a blended instructional approach for school *k*;

β_{1k} = the average adjusted log-odds associated with the mean response on covariate 1 (COV1) in school *k*;

β_{2k} = the average adjusted log-odds associated with the mean response on covariate 2 (COV2) in school *k*;

β_{19k} = the average adjusted log-odds associated with the mean response on covariate 19 (COV19) in school *k*. Note that at most 19 covariates could be included in the propensity model, however, no more than 7 classroom level covariates were included in the final subset of observed pre-treatment covariates per grade.

The level 2 model accounts for variation in the log-odds between schools on blended instructional approach. Therefore, the level 2 equation is

$$\pi_{0k} = \gamma_{00} + \gamma_{01}(SCOV1)_k + \gamma_{02}(SCOV2)_k + ... + \gamma_{019}(SCOV19)_k + u_{0k}, \quad (4)$$

$\pi_{nj} = \gamma_{n0}$ for n=1 to 19

where

γ_{00} is the average adjusted log-odds across schools in emphasizing a blended instructional approach;

γ_{01} = the average adjustment in log-odds associated with school covariate 1 (SCOV1) on emphasizing a blended instructional approach;

γ_{02} = the average adjustment in log-odds associated with school covariate 2 (SCOV2) on emphasizing a blended instructional approach;

γ_{019} = the average adjustment in log-odds associated with school covariate 19 (SCOV19) on emphasizing a blended instructional approach;

u_{0j} = the error associated with school *k*; and

γ_{n0} for n=1 to 19 are the average log-odds across schools for the respective 19 classroom level covariates in emphasizing a blended instructional approach. Note that at most 19 covariates could be included in the propensity models, however, no more than 8 school level covariates were included in the final subset of observed pre-treatment covariates per grade.

Subclassification of the propensity score. I use propensity score stratification (Rubin, 1997) to stratify the estimated logits derived from the propensity score models described above. Using this method, I first rank the estimated logits, divide the rankings into five percentile groups while disregarding assignment to treatment group, and then assign members of the five percentile groups to their respective strata such that five binary variables are created.

Propensity score theory claims that if the propensity scores are relatively constant within each subclass, then within each subclass, the distribution of all covariates should be approximately the same in both treatment groups. I test this property by accessing balance between treatment groups on each covariate while controlling for strata. To test balance, I use logistic regression. The outcome again is treatment or the

instructional approach. To the unconditional model, I add each pre-treatment covariate one by one, and I obtain the t-statistic representing its relationship with the treatment or outcome. I repeat these models including four of the five strata with one pre-treatment covariate at a time and again obtain the t-statistic. Significant t-statistics suggest a lack a balance. I present the results from the balance tests in Appendix C. After controlling for strata (provided in the conditional t-statistics column) using a linear regression model, balance was achieved between the two treatment groups on nearly all pretreatment covariates. Significant differences at the .05 level were found for (a) second grade imputed data set 3, variables mathematics and mathematics methods courses and professional development, (b) second grade imputed data set 5 variable professional development, and third grade imputed data set 1 variable courses.

Raw Achievement Differences between Common and Blended Classes

In this section, I compare the raw multi-digit computation scores for classes receiving common and blended approaches. Average fall multi-digit scores were higher for common second, third, and fourth grade classes but lower for fifth grade classes compared to blended classes. Figures III.1 and III.2 give the Fall and Spring logit scores for the multi-digit scale. By Spring, only second and fourth common classes scored higher than their blended peers, on average. For both Fall and

Spring assessments, average fifth grade blended classes outperformed their peers. Compared to the fall assessment, average third grade blended classes outpaced average common classes. Overall, this does not suggest that the blended instructional approach is superior to the common approach.

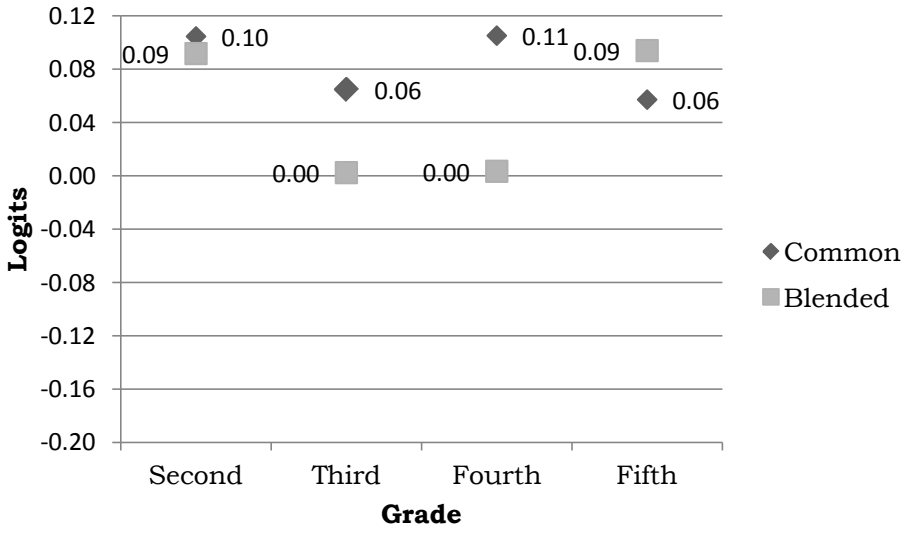

Figure III.1 Fall multi-digit scale scores, in logits, by grade and treatment groups

Recall that Fall and Spring multi-digit scale scores are not comparable. In order to get a sense of how much real mathematics learning occurred during the year between the treatment groups, I explore yearly gains using the TerraNova mathematics scale which are comparable. I used students Fall and Spring scores to estimate their yearly gain in mathematics. The results are similar to those for the multi-

digit scales. For second grade, the average gain in mathematics was about 43 points for common and 40 points for blended classes. For third grade, the average gain was 25 points for common and 29 points for blended classes. For fourth and fifth grades, the difference in gains between the treatment groups is smaller. At both grades, common classes on average gained about one point more than the average gain for blended classes. Again, this unadjusted comparison does not give evidence that the blended instructional approach yields greater gains in mathematics learning. However, given the differences in initial status, I can not draw causal conclusion about effect from instructional approach.

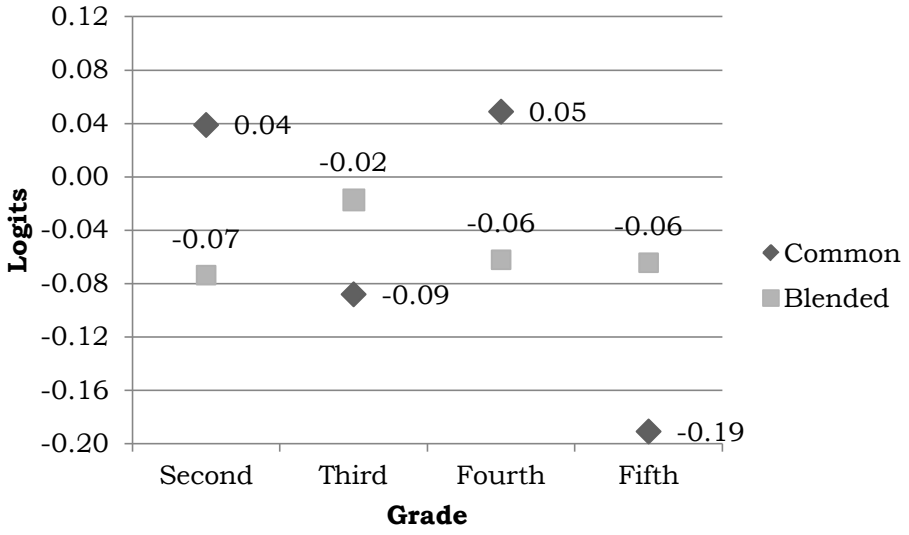

Figure III.2 Spring multi-digit scale scores, in logits, by grade and treatment groups

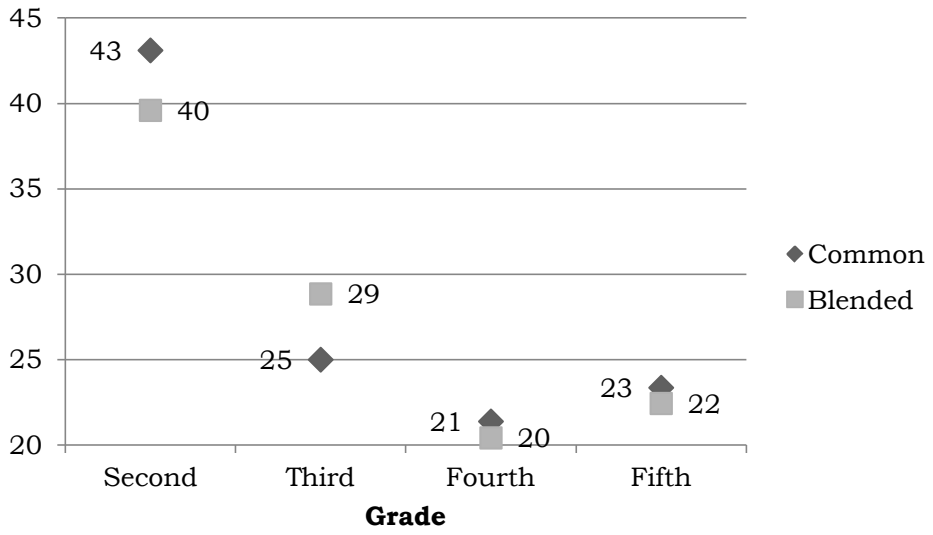

Figure III.3 Gain scores on TerraNova mathematics scale, by grade and treatment groups

Given the previous raw findings, I explore the within-stratum mean differences between common and blended classes. Within-stratum groups should be more similar on pretreatment variables. Table III.14 and Table III.15 give within-stratum descriptive statistics for common and blended classes and mean differences between treatment groups for each grade. Mean differences represent mean for common classes minus the mean for blended classes. Therefore, a positive difference suggests higher level of achievement for common classes and a negative difference suggests higher level for blended classes. For the Fall assessment and for all grades, at least three of the five strata showed positive difference. In

more strata, blended classes showed a lower level of mathematics achievement compared to common classes.

Results for Spring assessment are different. Again, a positive difference suggests a higher level of achievement for common classes and negative differences suggest a higher achievement for blended classes. For second grade, all within-stratum differences are positive and common classes showed a higher level of achievement compared to peers in blended classes. For third grade, all within-stratum differences are negative and blended classes showed a higher level of achievement. For both fourth and fifth grades, within-stratum differences are mixed but more differences are positive. In the upper grades, more strata suggest that common classes showed a higher level of achievement as compared to blended classes.

Table III.14

Descriptive statistics for Fall multi-digit scale score for classes receiving common and blended instructional approach, by grade

	Common			Blended			Mean
Stratum	N	Mean	S.D.	N	Mean	S.D.	Difference
Second Grade							
1	170	0.08	0.69	8	0.09	0.49	-0.01
2	124	0.24	0.65	64	-0.16	0.81	0.40***
3	88	0.00	0.74	103	-0.01	0.76	0.01
4	20	-0.02	0.66	159	0.18	0.66	-0.20
5	8	0.28	0.97	188	0.10	0.71	0.18
Third Grade							
1	133	0.17	0.60	41	-0.12	0.62	0.29**
2	121	-0.03	0.65	62	-0.05	0.66	0.02
3	76	-0.13	0.71	122	-0.04	0.64	-0.09
4	68	0.04	0.73	112	0.04	0.65	0.00
5	42	0.04	0.67	137	0.09	0.62	-0.05
Fourth Grade							
1	132	0.01	0.64	35	0.27	0.57	-0.26*
2	72	0.01	0.72	84	0.18	0.46	-0.17
3	29	0.00	0.62	106	-0.12	0.68	0.12
4	18	0.41	0.52	119	0.08	0.66	0.33*
5	12	0.23	0.73	163	-0.03	0.70	0.26
Fifth Grade							
1	138	0.08	0.68	3	-0.44	0.89	0.52
2	62	0.14	0.70	103	0.05	0.79	0.09
3	29	0.20	0.75	137	0.16	0.70	0.04
4	12	0.16	0.71	121	0.17	0.66	-0.01
5	6	0.21	0.56	129	-0.01	0.76	0.22

†p < .1, *p < .05, **p < .01, ***p < .001

Table III.15

Descriptive statistics for Spring multi-digit scale score for classes receiving common and blended instructional approach, by grade

	Common			Blended			Mean
Stratum	N	Mean	S.D.	N	Mean	S.D.	Difference
Second Grade							
1	170	0.02	0.78	8	-0.24	1.15	0.26
2	124	0.09	0.68	64	-0.06	0.75	0.15
3	88	0.02	0.81	103	-0.21	0.81	0.23*
4	20	0.29	0.80	159	0.07	0.74	0.22
5	8	0.26	0.86	188	-0.10	0.78	0.36
Third Grade							
1	133	-0.08	0.75	41	0.01	0.63	-0.09
2	121	-0.13	0.75	62	-0.12	0.71	-0.01
3	76	-0.15	0.69	122	0.01	0.65	-0.16
4	68	-0.13	0.74	112	-0.10	0.74	-0.03
5	42	-0.12	0.67	137	0.05	0.69	-0.17
Fourth Grade							
1	132	-0.08	0.74	35	0.05	0.65	-0.13
2	72	-0.09	0.76	84	0.09	0.65	-0.18
3	29	-0.01	0.84	106	-0.23	0.78	0.22
4	18	0.22	0.70	119	-0.05	0.67	0.27
5	12	0.40	0.71	163	-0.04	0.73	0.44*
Fifth Grade							
1	138	-0.05	0.80	3	-0.03	0.63	-0.02
2	62	-0.36	0.83	103	-0.22	0.81	-0.14
3	29	0.04	0.91	137	-0.05	0.75	0.09
4	12	-0.11	0.77	121	-0.03	0.82	-0.08
5	6	-0.13	0.33	129	-0.07	0.82	-0.06

†p < .1, *p < .05, **p < .01, ***p < .001

Statistical Causal Models

Now that the propensity score stratification is complete, I am now ready to set up the causal models that test the causal effect of the common instructional approach versus the blended instructional approach, the control and case groups, respectively, on student achievement. Here, I use a set of causal models that includes the propensity score stratification, as discussed previously, to control for possible confounding variables. Furthermore, I control for three levels of student achievement, or prior knowledge, on multi-digit computation of the fall assessment period of the TerraNova.

Research suggests that students' prior knowledge may influence both which instructional approach students receive and the effectiveness of the approach received. Furthermore, it suggests that students with similar abilities or prior knowledge are often tracked into the same class (Barr, Dreeben, & with Wiratchai, 1983; Rist, 1970a, 1970b; Rosenthal & Jacobson, 1968). Many, including teachers, believe that the two instructional approaches under study here are more favorable or appropriate for use with students of certain abilities (Gamoran, Nystrand, Berends, & LePore, 1995; Gamoran & Weinstein, 1998; Oakes, 1985, 1992). Imagine, for example, students with low prior knowledge are more likely to be tracked into classes which are more likely to receive, say, the common approach, while student with high prior knowledge are more likely to be tracked into classes receiving the blended approach.

First, analytic models lacking controls for prior knowledge could lead to inaccurate conclusions that one approach is more effective when really the effect was due to prior knowledge and selection bias. Alternatively, if I control for prior knowledge using the "raw" logit scores or the continuous variable, I may over estimate or under estimate the effect because group membership effects which treatment is given. Say, for example, that by chance my data is completely divided. By this I mean that all classes with low prior knowledge on average received the common approach while all classes with high prior on average received the blended approach. Further suppose that classes with mean prior knowledge on average were split 50/50 for receiving either approach. In the end, results in such a case are likely misleading if in practice students in each prior knowledge group are not equally likely to receive either instructional approach. Therefore, I device a causal model that estimates, not only central tendencies of instructional effects, but also the variation in the effects across relevant student groups.

Other realities may also affect the results of models lacking controls for prior knowledge. For example, students with high prior knowledge may do well no matter what, and students with low prior knowledge may do poorly with either approach. Likewise, the reverse argument may be true. Another concern with such a model is that gains by either group of students under either instructional approach may not be uniform. For example, one group may result in large gains in

achievement under said approach while the other group may show low gains. Therefore, ascertaining valid findings is only possible with studying or modeling group membership. These control variables, described in detail in the following paragraph, then allow me to assess treatment effects for each group and the corresponding hypotheses.

In this section I describe the specifics of the models used to estimate the effect of instructional approach on student achievement and the relevant student groups. Specifically, the model is a three-level hierarchical linear model with a student level, a classroom level, and a school level. At level 1, the student level, the model is a simple linear regression with indicator variables that control for prior achievement level on multi-digit computation (J. B. Carroll, 1963). Here, I control for prior achievement using the multi-digit scale scores from the fall assessments. Specifically, I have divided these fall achievement scores into three groups, or thirds. I use thirds as the resulting groups largely represent the student groups tied to arguments on appropriate instructional approach. I characterize the lowest third by the dummy or indicator variable FAL1. Here, FAL1 = 1 if child's fall achievement score on the multi-digit computation scale falls in the lowest third, otherwise, FAL1 = 0. Similarly, FAL2 = 1 if child's fall achievement score on the multi-digit computation scale falls in the middle third, otherwise FAL2 = 0. And likewise, FAL3 = 1 if child's fall achievement score on the multi-digit computation scale falls in the highest third, otherwise, FAL3 = 0. By

partitioning fall achievement into these three groups, I am able to study, at level 2, the effectiveness of the blended instructional approach for each group and whether that group membership plays a role in its effectiveness. I include only two of these variables as one has been left out as the reference group, namely the middle group, called middle Fall achievement group or FAL2. Furthermore, in the preceding model, I grand-mean center FAL1 and FAL3. This centering does not affect the level-1 estimates since the proportion of students in FAL1, FAL2, and FAL3 should be equal due to the way the variables were constructed. Grand mean centering, however, affects the interpretation of π_{0jk} and the corresponding level-2 model. For the un-centered case, π_{0jk} is the average spring achievement score on the multi-digit computation scale for students in FAL2 from classroom j in school k. It is possible that for some classrooms and schools this value does not exist because these classrooms and schools have no students in the FAL2 achievement category. Conversely, all classrooms and schools have an average achievement score. This is the primary reason I choice to grand-mean center the level-1 variables, FAL1 and FAL3.[xxii]

Finally, the level 1 equation is:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(FAL1)_{ijk} + \pi_{2jk}(FAL3)_{ijk} + e_{ijk} \quad (5)$$

where

Y_{ijk} is the Spring achievement score on the multi-digit computation scale for a level of the TerraNova for student *i* in classroom *j* in school *k*;

π_{0jk} = the average Spring achievement score on the multi-digit computation scale for classroom *j* in school *k;*

π_{1jk} = the average Spring achievement score on the multi-digit computation scale for students in lowest third on Fall achievement (FAL1) for classroom *j* in school *k;*

π_{2jk} = the average in Spring achievement score on the multi-digit computation scale for students in upper third on Fall achievement (FAL3) for classroom *j* in school *k;*

e_{ijk} = the error term, the difference of a students' Spring achievement score from the classroom average score for their Fall reference group, which is assumed to be normally distributed with a mean of zero and a standard deviation of σ^2.

At level 2, the classroom level, this is where I study the effects on classroom average achievement of common versus blended instructional approach on classroom average achievement, the main focus of this study. The two instructional approaches have fundamental characteristics that make them distinct from each other and yet uniform within the approaches. It is these distinct that support assumptions of homogeneity. [xxiii] From day-to-day or unit-to-unit there may be

differences in instructional approach but, annually, these differences are minimal, especially in their effect on "annual" student achievement.

The first equation depicts the average Spring achievement for the three Fall achievement groups as a function of the two instructional treatments, common and blended, and strata group membership. Blended = 1 if class received blended instructional approach and Blended = 0 if class received common instructional approach. Each stratum variable is coded a 1 if class in the stratum and 0 otherwise. The next two equations model the Spring achievement for students in the lowest and highest thirds on the Fall assessment as a function of the instructional treatment. All *Blended* and *Strata* variables were grand-mean centered. For the strata variables, there are essentially equal numbers of classes in each strata so entering these variables has little effect on the interpretation of β_{02k} through β_{05k}. Doing so, however, makes the interpretation of β_{00k}, β_{10k}, and β_{20k} clearer, especially since variable *Blended* is also grand-mean centered. Here, each $(Blended)_{jk}$, a dummy or indicator variable, is adjusted by the grand-mean or $(\overline{Blended})$.. which affects the interpretation of β_{00k}, β_{10k}, and β_{20k}. Each β_{00k}, β_{10k}, and β_{20k} is adjustment for differences among schools in the proportion of common classes and their interpretation reflect this adjustment.

Therefore, the level 2 equations are:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(Blended)_{jk} + \beta_{02k}(Stratum_1)_{jk} + \beta_{03k}(Stratum_2)_{jk} + \beta_{04k}(Stratum_3)_{jk} + \beta_{05k}(Stratum_4)_{jk} + r_{0jk} \quad (6)$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}(Blended)_{jk}$$

$$\pi_{2jk} = \beta_{20k} + \beta_{21k}(Blended)_{jk}$$

where

β_{00k} is the average Spring achievement score on the multi-digit computation scale for classrooms in school k regardless of instructional approach or stratum membership;

β_{01k} = the adjustment in class average Spring student achievement on the multi-digit computation scale with a blended instructional approach regardless of stratum membership;

β_{02k} through β_{05k} = the adjustment for being a member of $Stratum_1$ through $Stratum_5$,respectively, on class average Spring achievement for the average Fall achievement group as compared to the reference stratum holding instructional approach constant for the respective Fall achievement group and h varies from 1 to number of strata determined by stratification (less one for the reference group);

r_{0jk} = the error term at level 2, or the difference between the estimated class average Spring achievement and the actual average class achievement for average Fall achievement group, which is assumed to be normally distributed and have a mean of zero and variation of ω;

and β_{10k} and β_{20k} = the average in Spring achievement scores on the multi-digit computation scale for the lowest and highest Fall achievement group, respectively, for a classrooms in school *k*, regardless of instructional approach in use;

and finally, β_{11k} and β_{21k} = the adjustment, or treatment effect, in class average Spring student achievement in being in the lowest or highest Fall achievement group, respectively, on the multi-digit computation scale with a blended instructional approach.

The level 3, or school level, equations are:

$$\beta_{00k} = \gamma_{000} + u_{00k}, \quad (7)$$

$$\beta_{01k} = \gamma_{010},$$

$$\beta_{0(h+1)k} = \gamma_{0h0},$$

and $\beta_{qpk} = \gamma_{qp0}$ for q=1 & 2 and p=0 & 1

where

γ_{000} is, for the Fall achievement level 3, the average, or grand mean, Spring achievement score across schools regardless of instructional approach in use; γ_{010} = for middle third Fall achievement group, the average adjustment on Spring achievement for a blended instructional approach across all schools; γ_{0h0} = for average Fall achievement group, the average adjustment or effect of stratum membership across all schools; u_{00k} = for this same middle Fall achievement group, the school

level error term, the difference between the school k mean Spring achievement and the grand mean Spring achievement for schools receiving blended instructional approach, which is assumed to be normally distributed with a mean of zero and a variance of τ;

and γ_{100} and γ_{200} = for the lowest and highest Fall achievement groups, respectively, the average, or grand mean, adjustment in Spring achievement adjustment score across schools, regardless of instructional approach in use;

γ_{110} and γ_{210} = for the respective Fall achievement level, the average adjustment on the adjusted Spring achievement for a blended instructional approach across all schools;

Through separate analyses, this model is used with data from grades 2 through 5 and the respective levels of the TerraNova, levels 12 through 15. Additionally, the Wald Test is used to test the significance of strata or group variables. To the causal model stated above, I include $\sum_{h=1}^{4} \beta_{0(h+5)k}(Stratum_h) * (Blended)_{jk}$ where $\beta_{0(h+5)k}$ and l varies from 1 to 4, the total number of strata less one for the reference group, represents the treatment effect for each strata. The multi-parameter Wald test, in this case, tests the following hypotheses:

$$H_0 : \beta_{0(1+5)k} = \beta_{0(2+5)k} = \ldots = \beta_{0(h+5)k} = 0,$$

where $\beta_{0(h+5)k}$ is the effect of $(Strata_h)*(Blended)_{jk}$ on the intercept. For this group of variables if the Wald Test or resulting chi-square statistic is significant with a p-value less than α=.05 then I will conclude that it is highly unlikely that the observed estimated for $\beta_{0(h+5)k}$ could have occurred under the hypothesis stated above. Furthermore, I will conclude that the observed estimates for $\beta_{0(h+5)k}$ are not zero and the variables for $(Strata_h)*(Blended)_{jk}$ should be included in the model. Conversely, if the chi-square statistic is not significant, having a p-value of α ≥ .05 then I will conclude that it is highly likely that these observed estimates for $\beta_{0(h+5)k}$ could have occurred under that hypothesis stated above. Furthermore, the variables for $(Strata_h)*(Blended)_{jk}$ can be omitted from the model. Adjustments to the model above were made according to the results of the Wald Test.

A second set of Wald Tests addresses one of my secondary research questions: Does the benefit of blended instruction vary by level of prior knowledge? Three Wald Tests address the following hypotheses:

$$H_0 : \beta_{01k} = \beta_{11k} = 0,$$
$$H_0 : \beta_{01k} = \beta_{21k} = 0,$$
$$H_0 : \beta_{11k} = \beta_{21k} = 0,$$

Where β_{01k} is the effect of *Blended* on the intercept, or average fall multi-digit computation achievement and β_{11k} and β_{21k} are the effects of *Blended* on the slope of the lowest and highest fall multi-digit computation

achievement groups, respectively. If the Wald Test or the resulting chi-squared statistic is significant using a cut-off p-value of α=.05, I will conclude, for the pairwise comparison, that it is highly unlikely that the observed estimates β_{01k} or β_{01k} or β_{01k} for the treatment effect is different for the two groups. Conversely, if the results are not significant, I will conclude, for the pairwise comparison, that the treatment effect is not different for the two groups. In this case, a model that only studies the effect of *Blended* on the average Fall multi-digit computation achievement is reasonable, however, the current model represented by equations 1-3 model current theory and a discussion of the results of the results is likely to be as informative to instructional practice and future research. Therefore, in either case, no changes will be made to the causal model.

In Chapter 4, I present the results from multiple imputation and the predictive and causal models.

Chapter IV

Results

In Chapter III, I presented the appropriate methods for addressing my research questions. The prominent research question addressed by this dissertation is "would the average class receiving a treatment of common instruction learn significantly more had they received blended instruction in the same school?" Furthermore, I explore answers to the question, Does the benefit of blended instruction vary by level of prior knowledge? Are there grade-level differences? And, what affects teachers' choices of instructional approach? Specifically, is there a relationship between teachers' knowledge, gender, and ethnicity; class average ethnicity, gender, and achievement; and school day length, minority composition, whole school reform participation, curriculum endorsement, average teacher content knowledge for teaching, and average achievement for entering kindergarteners and the enacted instructional approach? Addressing such questions about instruction is new territory. In this chapter, I present the results of my inquiry.

I divide the chapter into three sections. In the first section, I present the results from multiple imputation for missing data. In the

next section, I present the results from the analytic models, first the predictive model results followed with the causal model results. In the final section I discuss the limitations.

Multiple Imputation

As is typical for survey data, data from the Study of Instructional Improvement (SII) used in this study contain observations with missing values on variables. Past methods for dealing with missing data usually resulted in statistical procedures which exclude from the analysis observations with any missing variable values or imputing known values such as the mean. These exclusionary methods also meant that valid values on variables were removed from the analyses. In cases when mean values are imputed, standard errors of the variables at issue are narrowed compared to the true standard errors, and results are, therefore, compromised. Multiple imputation (MI) methods address both of these limitations. They eliminate the need for exclusion of observations with missing variable values by replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Furthermore, the process of choosing plausible values also strives to preserve the original mean and standard errors. MI also has the advantage of preserving all valid responses on variables. Given the large amount of missing data and the positive attributes of MI, I chose to use it in this analysis and begin this chapter by presenting those results.

Because the imputed values were needed prior to statistical modeling and the integrity of the imputed data impact the results from the causal and predictive models, I present the results here and not in the previous chapter, the methods chapter.

Imputation procedures are discussed in Chapter III, Methods, but recall that, for imputation purposes, I use all available and pertinent data from SII which could potentially that could potentially inform the imputation model and improve accuracy and reduces bias in the imputed data. Therefore, I used more data than I expected to use in the statistical models.

The results from multiple imputation, a procedure implemented in MICE 2.0, are provided in Appendix B. For details about MICE, see Chapter III, Methods, where I discuss the imputation method used and its limitations.

MICE assumes that missing data are missing at random (MAR). Departures from MAR can arise when large fractions of data are missing as is the case in the SII data. There is currently no test to assess whether MAR is violated (Potthoff, Tudor, Pieper, & Hasselblad, 2006). However, there may be evidence that assuming MAR is or is not appropriate from studying patterns of missing data.

In terms of the SII data, many variables are missing many data points, and, for many cases, many variables are missing data. A large portion of the data heavily used in this study, came from annual

Teachers Questionnaire which was administered every year to every teacher in the SII study. Some missing data is due to not returning the questionnaire. In other cases, since the questionnaire was quit long, a bulk of the missing data came at the end of the questionnaire. This type of missing pattern meant that the later section(s) contained items with missing responses more often. The last sections asked about demographic, credential and training. By looking at available data, I could assess whether responses are missing because respondents did not want to give particular answers. It does not seem that the range of data was affected and therefore, missing data are unrelated to the response. This seemed to be true across all items. That is, the expected range of responses was evident in the available data. Furthermore, for some questionnaire items, the instrument was given many times over the four year, therefore, in the case of missing data, I could extrapolate and estimate of the true response. Here again, there are no patterns in the missing data. Therefore, without a specific test for MAR, the evidence as it exists supports assuming MAR and using multiple imputation.

In Appendix B, I report, for comparison and assessment of the quality of the imputed data, descriptive statistics for the complete data and for the five multiply imputation data sets. Since the classroom level variables are school-mean centered for the imputation procedure, this table provides the descriptive statistics of centered data. The data are then un-centered for use in the statistical procedures. The means and

standard deviations of the original, complete data and the five combined, complete plus imputed, data should be approximately the same if the imputation routine successfully imputed plausible values which I find to be true. A check for senseless values is also warranted.

For each variable where data was imputed, I first checked to see that the imputed data falls within the minimum and maximum, imputing no extreme or outlying values. Next, I compared the means and standard deviations. Since the original data are group mean centered and therefore have a mean of zero, I checked the imputed data which also have means reasonably close to zero. Therefore, from the imputation results, this all shows that the imputation procedure in MICE 2.0 was successful.

With the successfully imputed data, I now precede with the results from the statistical models.

Analytic Models

At this point in the analysis, I begin the model building that directly addresses my research questions. Here, I move from using all available and relevant SII data and results from multiple imputation to only data useful to the grade specific analysis. Results are presented for each grade. I begin with results for the predictive models.

Predictive Models

What influences a class receiving a blended instructional approach? In Chapter III, Methods, for each grade, I propose a predictive model containing seven classroom-level covariates and six school-level covariates to address this question. I use a 2-level logistic regression with these covariates modeling the intercept. See Equations 6 and 7 in Chapter III. This model assesses the relationship between instructional approach and class composition on gender, ethnicity, and achievement and teacher gender, ethnicity, and content knowledge for teaching. It also accesses the relationship between instructional approach and school characteristics and aggregate student and teacher characteristics.

Second grade results. I present the results in grade order beginning with second grade. See

Table IV.1 and Table IV.2 for results from unconditional and conditional models, respectively. The average second grade class in average schools has a 59 percent chance of receiving blended instruction. There is only a moderate reliability (r = .41) among schools to distinguish on receiving a blended instructional approach. Only a few variables show a significant association with the treatment assignment, blended instruction. For second grade, there are no classroom-level variables which show a significant association with receiving blended instruction. That is, these results suggest that class gender, ethnicity, and prior knowledge and teachers' gender, minority status, and content

knowledge for teaching are unrelated to second grade classes receiving blended instructional approach.

There are, conversely, several school-level variables which show a significant association. First, school minority composition (percent African American (AAS)) and involvement in whole school reform (ACSFA) are unrelated to second grade class's probability in receiving blended instructional approach. The coefficients for length of school day (LENGTH) and school average Woodcock Johnson mathematics score (WOOD) are negative and significantly associated with receiving blended instruction, coefficients -1.85 (SE = 0.9) and -0.40 (SE = 0.16), respectively. Recall that WOOD represents the mathematics achievement score for kindergarteners. Thus, classes in schools of average LENGTH and average WOOD have a reduced probability in receiving blended instruction, and schools having WOOD or LENGTH one or even two standard deviations greater are even more likely to not receive blended instruction. This last finding is consistent with research findings suggesting that students in low academic tracks or groups receive instruction that is less cognitively demanding. Furthermore, given these school level achievement results and the fact that class characteristics are unrelated to the instructional approach received, these results suggest that second grade teachers' perceptions have more influence on the instructional approach they support than the characteristics of the

students they were teaching. This conclusion is also supported by research on tracking and instructional differentiation.

Table IV.1
Results from unconditional 2-level logistic regression model, second through fifth grades

	Fixed Effect, γ_{00} (SD[a])	Probability, $\frac{1}{1+e^{-\gamma_{00}}}$	Among Schools, μ_{00}	
			Variance	Reliability
Second grade (n= 180 classrooms)	0.38 (0.20)	0.59	1.35***	.41
Third grade (n=171 classrooms)	0.06 (0.15)	0.51	0.01	.00
Fourth grade (n=171 classrooms)	0.80 (0.19)	0.69	0.87**	.29
Fifth grade (n=158 classrooms)	0.82 (0.23)	0.69	2.22***	.50

*** p <.001, ** p <.01, * p <.05
[a] SD = standard deviation

The coefficients for school average teachers' content knowledge for teaching (AVEKNOW) and NSF curriculum in use (NSF) are positive and significantly related to receiving blended instructional approach, namely 2.21 (SE = 1.10) and 1.22 (SE = 0.58), respectively. Furthermore, these coefficients are large compared to the other coefficients in the model. These results suggest that having highly competent teacher colleagues and an NSF curriculum in use at the school that supports or models a blended instructional approach improves markedly second graders probability in receiving blended instruction.

The second grade model supporting these results written in mixed form is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = 0.47 + 1.11(SMALE) - 0.25(SAA) + 0.92(SHISP) + 0.09(ACHIEVE)$$
$$+ 0.35(TMALE) - 0.03(TCAUC) + 0.26(KNOW) - 1.85(LENGTH) +$$
$$0.00(AAS) + 2.21(AVEKNOW) - 0.40(WOOD) + 0.37(ACSFA) + 1.22(NSF) + 0.95.$$

Third grade results. Third grade results are similar to second grade results. The average third grade class in average schools has a 51 percent or nearly equal chance of receiving blended instruction versus common instruction. There is very low reliability (r = .00) among schools to distinguish on receiving a blended instructional approach, so I cautiously report these results. Here again, only a few variables show a significant association with the treatment, blended instruction. Similar to second grade results, there are no classroom-level variables which show a significant association with receiving blended instruction in third grade. Again, these results suggest that class gender, ethnicity, and prior knowledge and teachers' gender, minority status, and content knowledge for teaching are unrelated to third grade classes receiving blended instructional approach.

There are three school-level variables which have a significant association and three covariates which have a non-significant relationship with third grade classes receiving blended instructional approach. First, school average teachers' content knowledge for teaching (AVEKNOW), involvement in school reform (ACSFA), and NSF curriculum

in use (NSF) are unrelated to third grade class's probability in receiving blended instruction. Conversely, length of school day (LENGTH), school minority composition (or percent of African American students [AAS]), and average kindergarten mathematics achievement (WOOD) are negatively and significantly related to third grade classes receiving blended instruction. The estimated effect for average kindergarten mathematics achievement (WOOD) is small at -0.28 (SE = 0.14) and for school minority composition (AAS) even negligible at -0.03 (SE = 0.01), but the estimated effect of length of school day (LENGTH) was comparatively large at -2.14 (SE = 0.63). Overall, these third grade results, like the results for second grade, suggest that factors influencing instructional approach lie outside the classroom.

Written in mixed form, the third grade model supporting these results is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = 0.04 + 0.04(SMALE) + 1.96(SAA) + 0.57(SHISP) + 0.43(ACHIEVE)$$
$$+ 0.40(TMALE) + 0.08(TCAUC) + 0.00(KNOW) - 2.14(LENGTH) - 0.03(AAS) +$$
$$1.56(AVEKNOW) - 0.28(WOOD) + 0.73(ACSFA) + 0.32(NSF) + 0.00.$$

Fourth and fifth grade results are markedly different from results for second and third grades. Average fourth and fifth grade classes in average schools have a 69 percent chance of receiving blended instructional approach. The reliability to distinguish between schools on their log-odds of receiving blended instruction in fourth grade is moderately low at .29 and for fifth grade, it is moderate at .50.

Results for fourth grade show significant relationship between classroom covariates and receiving blended instruction but not between school covariates and receiving blended instruction. Class minority composition (proportion African American (SAA) and Hispanic (SHISP)), average class prior achievement (ACHIEVE), and teachers' content knowledge for teaching (KNOW) are not significantly related to receiving blended instruction. Conversely, students' gender (SMALE) and teacher's gender (TMALE) are positive and significantly related, while teacher's ethnicity (TCAUC) is negative and significantly related. The estimated effects of students' gender (SMALE), teacher's gender (TMALE), and teacher's ethnicity (TCAUC) are 2.92 (SE = 1.09), 1.74 (SE = 0.71), and -1.02 (0.52), respectively. Thus, fourth grade classes having higher proportion of males and a non-Caucasian male teacher are more likely to receive blended instruction. In contrast to second and third grade results, length of school day (LENGTH), average student characteristics (AAS and WOOD), school resources (AVEKNOW and NSF), and reform program endorsement (ACSFA) are not influential on instructional approach. Alternatively, fourth grade instructional approach is influenced, not by external factors as is the case with second and third grades, instead by internal factors such as teacher and student characteristics. These results could be viewed also as support for claims from research on tracking that instructional difference are due to teachers' perceptions. Given the lack of influence from average class

prior achievement and teachers' content knowledge of teaching, the significant covariates suggest that instruction in fourth grade class is influenced by cultural biases. There are long standing stereotypes that boys do better in mathematics than girls and are more likely to be asked more demanding questions. From these results, classes with more males are more likely to receive an instructional approach that emphasizes greater cognitive demanding, here the blended instructional approach. The positive significant effect of male teachers and the negative significant effect of Caucasian teachers on instructional treatment are less clear but maybe related to instructional biases.

The fourth grade model supporting these results written in mixed form is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = 1.10 + 2.92(SMALE) - 0.37(SAA) - 0.53(SHISP) - 0.27(ACHIEVE) + 1.74(TMALE) - 1.02(TCAUC) + 0.09(KNOW) - 0.78(LENGTH) - 0.01(AAS) - 0.98(AVEKNOW) + 0.15(WOOD) - 0.28(ACSFA) + 0.2(NSF) + 2.61.$$

Fifth grade results. Results for fifth grade are inconsistent with those for the previously reported grades. Of the six classroom-level and seven school-level independent variables, none of them are significantly related to receiving blended instructional approach using a cut-off p-value of .05. Proportion of Hispanic students is the only variable that has a p-value less than .10 which has an estimated coefficient of 2.82 (SE = 1.59). This model, containing 13 covariates, may over-fit the fifth grade data, but I report the results for this model so that grade comparisons

can be made. As is, these results suggest that class, teacher, and school characteristics do not influence instructional approach in fifth grade.

Results from a very parsimonious model containing only two classroom-level covariates, namely, proportion of Hispanic students (SHISP) and teachers content knowledge of teaching (KNOW), suggest a different story. See Table IV.3 for these results. The magnitude of the coefficients for these two variables are similar in the two models, the "13-covariate model" and the "2-covaraite model," which further suggests that the model containing 13 covariates over-fits the data. Results from the two-covariate model suggest that proportion of Hispanic students (SHISP, coefficient = 2.13, SE = 0.89) and teacher's content knowledge for teaching (KNOW, coefficient = 0.48, SE = 0.23), controlling for no other factors, are positive and significantly related to fifth grade classes receiving a blended instructional approach. That is, the greater the proportion of Hispanic students in the class and the greater the teacher's content knowledge for teaching the more likely the class will receive a blended instructional approach. These findings are consistent with recommendation from the *Standards*, recommendation that support a blended instructional approach for all students regardless of their ethnicity. Unlike results for other grades, these findings are not consistent with research on tracking. This area of research finds that students with a less privileged social background receive instruction that is less demanding, on the one hand, but there are also findings that

suggest the teachers' make instructional choices in response to their students needs. How these findings relate to the fifth grade data is less clear. Both measures of social background and class average achievement are not related to instructional approach. Exploring interaction terms might explain the relationship between proportion Hispanic and teacher's content knowledge for teaching with instructional approach. This, however, is left for further research.

The fifth grade model supporting these results written in mixed form is

$$\log\left(\frac{\mu_{jk}}{1-\mu_{jk}}\right) = 1.14 + 1.32(SMALE) - 3.00(SAA) + 2.82(SHISP) - 0.27(ACHIEVE)$$
$$+ 0.52(TMALE) + 0.62(TCAUC) + 0.48(KNOW) - 1.31(LENGTH) +$$
$$0.03(AAS) - 0.79(AVEKNOW) + 0.03(WOOD) - 0.03(ACSFA) + 0.69(NSF) + 4.83.$$

Across grade comparisons. There are across grade-level similarities and differences. Second and third grade results are more alike while results for fourth and fifth grades share some commonalities. In the lower grades, second and third, instructional influences are external. On average, instruction on multi-digit computation begins in second grade where they primarily focus on multi-digit addition and incorporate subtraction later in the year. Multiplication and division are new and the center of the introductory work in third grade. In light of the content focus, it seems from these results that teachers find support in their instructional choices from printed resources and colleagues and their perceptions of students needs. These results may also suggest that

instructional approach is planned in advance, and, therefore, classroom characteristics, internal influences, are inconsequential.

The story changes in fourth and fifth grades when the focus is less on introducing new material in terms of new computational procedures. In these grades, the content is a mix of new procedures that require competencies on earlier computation procedures. For example, long division is new content, having specific procedural steps and concepts. Success, however, relies on knowledge of basic multi-digit computation skills, and students' prerequisite knowledge is more likely to influence instruction.

This research found that school characteristics have no influence on instructional approach in the upper grades while classroom characteristics do. The specific influences are mixed. In fourth grade, stereotypes affect instructional approach but not in fifth grade. One explanation might be that teachers lack "real" measures of their prerequisite knowledge and stereotypical expectation fill in the missing information. Furthermore, as teachers work with their students they identify information that affirms the stereotypes and ignores contrary evidence. In fifth grade, while the results are questionable, there is evidence that professional guidance plays a role. This is sensible since fifth grade teachers are responsible for getting students ready for sixth grade which is often a middle school grade. In sixth grade, students often "change classes" and have subject specific teachers who may

communicate expected prior knowledge for entering sixth graders in an effort to coordinate multiple schools feeding into the middle school.

In summary, for both second and third grades, the estimates for length of school are one of the largest and significant, but not the case for fourth and fifth grades. Comparatively, long days for younger elementary classrooms contributes to less demanding instructional choices, while it seems inconsequential for older elementary classrooms. Colloquial support seems to help combat this. On the other hand, the strength of gender stereotypes seems to surface in the later graders.

Further studies are needed to make sense of these results. Overall, the results seem sensible and one explanation suggests that there is complicated interplay between instructional approach and content, prerequisite knowledge, stereotypes, teacher knowledge, and curriculum.

Table IV.2
Results of conditional 2-level logistic regression models, second through fifth grades

Predictor	Grade 2 B (SE)	Grade 3 B (SE)	Grade 4 B (SE)	Grade 5 B (SE)
Intercept	0.47*	0.04	1.10***	1.14*
	(0.21)	(0.19)	(0.26)	(0.35)
Classroom Level				
Student Characteristics				
Proportion males (SMALE)	1.11	0.40	2.92**	1.32
	(0.98)	(1.00)	(1.09)	(1.21)
Proportion African American (SAA)	-0.25	1.96†	-0.37	-3.00
	(1.18)	(1.17)	(1.59)	(1.90)
Proportion Hispanic (SHISP)	0.92	0.57	-0.53	2.82†
	(0.88)	(0.85)	(1.18)	(1.59)
First principal component of average achievement on Reading, Language arts, & Mathematics (ACHIEVE)	0.09	0.43	-0.27	-0.27
	(0.21)	(0.29)	(0.31)	(0.71)
Teacher Characteristics				
Male teacher (TMALE)	0.35	0.40	1.74*	0.52
	(0.71)	(0.77)	(0.71)	(0.78)
Teacher is Caucasian (TCAUC)	-0.03	0.08	-1.02*	0.62
	(0.42)	(0.43)	(0.52)	(0.52)
Measure of Content Knowledge for Teaching (KNOW)	0.26	0.00	0.09	0.48
	(0.22)	(0.22)	(0.23)	(0.41)
School Level				
Length of school day (LENGTH)	-1.85**	-2.14***	-0.78	-1.31
	(0.69)	(0.63)	(0.83)	(1.10)
Percent African American Students (AAS)	0.00	-0.03*	-0.01	0.03
	(0.01)	(0.01)	(0.02)	(0.02)
Teacher average Content Knowledge for Teaching (AVEKNOW)	2.21*	1.56	-0.98	-0.79
	(1.10)	(0.98)	(1.46)	(1.76)
Average Woodcock Johnson Mathematics score (WOOD)	-0.40*	-0.28*	0.15	0.03
	(0.16)	(0.14)	(0.17)	(0.24)
American's Choice or Success for All School (ACSFA)	0.37	0.73†	-0.28	-0.03
	(0.47)	(0.38)	(0.55)	(0.07)
NSF curriculum in use (NSF)	1.22*	0.32	0.62	0.69
	(0.58)	(0.48)	(0.67)	(0.85)
Random Effect (Variance)				
Between-school, u_{0j}	0.95*	0.00	2.61***	4.83***

†p < .10, *p < .05, **p < .01, ***p < .001

Table IV.3
Results for conditional model at level 1 and unconditional at level 2, fifth grade

Predictor	Grade 5 B (SE)
Intercept	0.92** (0.25)
Classroom Level	
Student Characteristics	
Proportion Hispanic	2.13* (0.89)
Teacher Characteristics	
Measure of Content Knowledge for Teaching	0.48* (0.23)
School Level	
-unconditional-	
Random Effect (Variance)	
Between-School, u_{0j}	2.46***

†p < .10, *p < .05, **p < .01, ***p < .001

Statistical Causal Models

Would the average class receiving a treatment of common instruction learn significantly more had they received blended instruction in the same school? In this section, I discuss the results of the causal models which answer this question. I define blended and common instructional approach similarly for all grades. Blended instructional approach is an instructional emphasis during a focus on multi-digit computation that incorporates a high emphasis on procedures and a moderate to high emphasis on concepts. Alternatively, common instructional approach is an instructional approach during a

focus on multi-digit computation that incorporates a high emphasis on procedures and no or low emphasis on concepts. I present the results for each grade beginning with second grade.

I estimated the results for the causal models using the 3-level HLM model described by Equations 5-7 given in Chapter III, Methods (Raudenbush, Bryk, & Congdon, 2009). Recall that I hypothesize that blended instruction is positively and significantly related to the average class Spring achievement score on the multi-digit computation scale. Prior to obtaining the results to the causal models, I begin with a statistical analysis of the data structure using an unconditional model. See Table IV.4 for the variance decomposition of the grade level measures of achievement on multi-digit computation. Between 80 to 91 percent of the variation in spring achievement on multi-digit computation lies between students, while 4.6 to 9.5 percent lies among classrooms within schools and 2.8 to 14.4 lies among schools. For second and fifth grades, there is more variation in spring achievement among schools than among classrooms within schools. Conversely, for third and fourth grades, there is more variation among classrooms within school than among schools. For all grades, the reliability to distinguish among classrooms within schools and among schools on spring multi-digit computation achievement is moderate to moderately low. The differences between grades are reasonable since the test items and the number of items used to measure achievement on multi-digit computation also differ by grade.

Using a statistical significance cut-off of α=.05, the variance among classrooms within schools was significant for second and third grades only. The variance, however, among schools was significant for all grades.

Table IV.4
Variance Decomposition of Achievement on Multi-digit Computation from the unconditional models, second through fifth grades

	Among Students within Classrooms, ε	Among Classrooms within Schools, ρ_0		Among Schools, μ_{00}	
	Percent Variance	Percent Variance	Reliability	Percent Variance	Reliability
Second grade (n=932 students)	80.0	9.5***	.365	10.5***	.431
Third grade (n=914 students)	90.0	7.3*	.294	2.8*	.167
Fourth grade (n=770 students)	90.9	4.6	.182	4.4**	.249
Fifth grade (n=740 students)	80.5	5.1	.222	14.4***	.518

*** p <.001, ** p <.01, * p <.05

I continued model development by fitting the models described by Equations 5-7 in Chapter III, Methods. Before reporting the results, I assessed whether the interaction terms describing strata by match group assignment, written as $\beta_{0(h+5)k}(Strata_h)*(Blended)_{jk}$, are needed in the

model. I used a multi-parameter Wald test, with results given in Table IV.5, to test whether the coefficients for the interaction terms are all equal to zero. If this test results in a significant p-value, $\alpha \leq .05$, then the treatment effect is not constant across groups and the interactions should stay in the model. Conversely, if this test results in a non-significant p-value, $\alpha > .05$, then the treatment effect is the same across the matched groups and the interaction terms can be removed from the model. For all grades, the Wald test results — chi-square statistics, degrees of freedom, and p-values — yielded p-values greater than a cut-off of .05, so, therefore, I failed to reject the null hypothesis and do not include the interaction terms in the final causal models.[xxiv]

Table IV.5
Results of Wald Test used with Causal Models, composite hypothesis tests of interaction of treatment by strata for second through fifth grades

	Chi-square Statistic	Degrees of Freedom	P-value
Second grade (n=932 students)	2.12	4	>.50
Third grade (n=914 students)	1.38	4	>.50
Fourth grade (n=770 students)	6.79	4	.15
Fifth grade (n=740 students)	1.24	4	>.50

*** p <.001, ** p <.01, * p <.05

Using the results given in Table IV.6 through Table IV.9, I assess the results from the causal models for statistical significance using a cut point of α=.05. These models test the causal effect of blended instruction on spring student achievement. Blended instruction, an annual measure

of instructional emphasize on conceptual and procedural knowledge during content focused on multi-digit computation, is compared to common instruction, an annual measure instruction emphasizing, primarily, procedural skills during work on multi-digit computation. These models also tested the effect of blended instruction on prior fall achievement on multi-digit computation. Recall that the hypotheses being tested are that blended instruction as compared to common instruction is (a) positively and significantly related to average classroom achievement on multi-digit computation and (b) more beneficial to high achieving students that lower achievement students, as measured in the fall. Results vary by grade.

For second and fourth grades, the estimate for the causal effect of blended instructional approach on multi-digit computation achievement was negative with coefficient estimate of -0.12 (standard error (SE) = 0.07) and -0.04 (SE = 0.07), respectively, and not statistically significant. Conversely, for third and fifth grades, the estimated blended instructional effect was positive at 0.09 (SE = 0.06) and 0.11 (SE = 0.09), respectively and also not statistically significant. Therefore, in terms of my two part hypothesis stated in the previous paragraph, part (a) is rejected. While no treatment effect was significant, the direction of the effect differed across the grades. For second and fourth grades, the effect is positive, while, the effect is negative for third and fifth grades. The mixed directional effects are puzzling. Recall that students who took

different test levels are not included in this analysis due to methodological limitations for generating scores that are comparable across forms. These students were taking more or less difficult test levels and may be the beneficiates of the blended instructional approach. In the end, by not including these students, I may have distorted the effectiveness of blended instructional approach. I am cautiously optimistic that when a full range of students are included blended instruction can have significant effects on learning.

Another part of the causal model estimates the blended instructional effect on the slopes of the lower and upper third achievement groups. Since the slope models the rate of change in achievement for each of these achievement groups, the coefficient of treatment in Equation 4 then represents the estimate of the difference in the rate of change in achievement in favor of blended versus common instruction. The results estimating the treatment effect on slope or rate are relatively consistent across the grade level models. For all grades, the blended instruction effect on lower third achievement group slope was negative and not statistically significant, with estimates between -0.10 (SE = 0.13) and -0.01 (SE = 0.13). The instructional effect of treatment on upper third achievement group slope was negative and not statistically significant for second and fifth grades, with estimates -0.08 (SE = 0.12) and -0.22 (SE = 0.13), respectively. Conversely, for third and fourth grades, the effect of treatment on upper third achievement group slope

was positive and not statistically significant with estimates 0.04 (SE = 0.11) and 0.06 (SE = 0.13), respectively. Therefore, given of the non-significant findings and negligible effects, it is highly unlikely that there is a significant difference in rate of change between classrooms who receive blended instruction compared to those who receive common instruction in both lower and upper achievement groups. Furthermore, these results provide no evidence that blended instruction is more or less beneficial given initial achievement status and therefore I reject my hypothesis that blended instruction is particularly beneficial to low achieving students.

Again, these results are limited by the pool of students' scores that are usable for this study. The student groups defined to be in the lower and upper third of achievement groups would likely not be in those groups if I was able to include all students assessed by SII. Recall that students who were expected to outperform or underperform on the standard test level were given an alternate level, and, for this study scoring methods limited my ability to include these students. As a results, I expect that this study excludes an unknown proportion of high and low achieving students, some of the who would have been included the high and low achievement groups defined in this study. Therefore, the non-significant blended instruction effect is not surprising when there was also a non-significant effect on average students.

Table IV.6
Results for second grade causal model: Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error
School Level – Intercept, γ_{000}	-0.02	0.03
Classroom Level – Intercept		
Treatment, β_{01k}	-0.12	0.07
Stratum 1, β_{02k}	0.06	0.10
Stratum 2, β_{03k}	0.12	0.09
Stratum 3	Reference group	---
Stratum 4, β_{04k}	0.14	0.09
Stratum 5, β_{05k}	0.04	0.10
Classroom Level – Slope		
Lower Third Achievement Group, β_{10k}	-0.42***	0.06
Treatment, β_{11k}	-0.10	0.13
Upper Third Achievement Group, β_{20k}	0.23***	0.06
Treatment, β_{21k}	-0.08	0.12

Variance Components	Estimate	Percent Variance Explained
Between Students within Classrooms, e_{ijk}	0.42	11.3
Between Classrooms within Schools, r_{ojk}	0.03**	50.0
Between Schools, u_{00k}	0.04***	40.0

***p<.001, **p<.01, *p<.05

Table IV.7

Results for Third grade causal model: Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error
School Level – Intercept, γ_{000}	-0.06*	0.03
Classroom Level – Intercept		
Treatment, β_{01k}	0.09	0.06
Stratum 1, β_{02k}	0.00	0.09
Stratum 2, β_{03k}	-0.05	0.09
Stratum 3	Reference Group	---
Stratum 4, β_{04k}	-0.07	0.09
Stratum 5, β_{05k}	0.02	0.09
Classroom Level		
Lower Third Achievement Group, β_{10k}	-0.19**	0.05
Treatment, β_{11k}	-0.08	0.11
Upper Third Achievement Group, β_{20k}	0.19**	0.06
Treatment, β_{21k}	0.04	0.11

Variance Components	Estimate	Percent Variance Explained
Between Students within Classrooms, e_{ijk}	0.44	3.4
Between Classrooms within Schools, r_{0jk}	0.03*	9.3
Between Schools, u_{00k}	0.01	55.8

***p<.001, **p<.01, *p<.05

Table IV.8
Results for Fourth grade causal model: Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error
School Level – Intercept, γ_{000}	-0.05	0.03
Classroom Level – Intercept		
Treatment, β_{01k}	-0.04	0.07
Stratum 1, β_{02k}	0.06	0.10
Stratum 2, β_{03k}	0.13	0.09
Stratum 3	Reference Group	---
Stratum 4, β_{04k}	0.10	0.10
Stratum 5, β_{05k}	0.14	0.09
Classroom Level – Slope		
Lower Third Achievement Group, β_{10k}	-0.23***	0.06
Treatment, β_{11k}	-0.01	0.13
Upper Third Achievement Group, β_{20k}	0.21**	0.06
Treatment, β_{21k}	0.06	0.13

Variance Components	Estimate	Percent Variance Explained
Between Students within Classrooms, e_{ijk}	0.46	4.6
Between Classrooms within Schools, r_{0jk}	0.02	27.9
Between Schools, u_{00k}	0.02*	25.3

***p<.001, **p<.01, *p<.05

Table IV.9
Results for Fifth grade causal model: Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard Error
School Level – Intercept, γ_{000}	-0.10*	0.04
Classroom Level – Intercept		
Treatment, β_{01k}	0.11	0.09
Stratum 1, β_{02k}	0.12	0.13
Stratum 2, β_{03k}	-0.15	0.10
Stratum 3	Reference Group	---
Stratum 4, β_{04k}	-0.06	0.10
Stratum 5, β_{05k}	0.02	0.10
Classroom Level – Slope		
Lower Third Achievement Group, β_{10k}	-0.48***	0.06
Treatment, β_{11k}	-0.05	0.14
Upper Third Achievement Group, β_{20k}	0.37***	0.06
Treatment, β_{21k}	-0.22	0.13

Variance Components	Estimate	Percent Variance Explained
Between Students within Classrooms, e_{ijk}	0.44	16.4
Between Classrooms within Schools, r_{0jk}	0.01	71.5
Between Schools, u_{00k}	0.06***	37.0

***p<.001, **p<.01, *p<.05

Using Wald Tests, I apply a second set of multi-parameter tests involving the slopes or rate in change in achievement. I examine whether a blended instructional approach has a similar effect on the intercept and slopes. Again, the intercept models the average effect while the slopes model the rate of change for particular groups, here the lower and upper achievement groups. These results are given in Table IV.10. The chi-square statistics range from 0.34 to 4.46 for the hypothesis tests

comparing a blended instruction effect on intercept and slope of both lower and upper third achievement groups and are not statistically significant. Therefore, it is highly unlikely that the blended instruction effect differs significantly for the intercepts and slopes for any grade. I also compared the blended instruction effect of blended instruction on the slopes of lower and upper third fall achievers. For all grades, there is no significant difference in this effect with Chi-squares ranging from 0.30 to >3.14. As previously mentioned, these results are not surprising since the range of students is limited due to methodology constraints on comparing students' achievement scores across different test levels.

Table IV.10
Results of Wald Test used with Causal Models, composite hypothesis tests of treatment effect on intercepts and slopes

	Chi-square Statistic	Degrees of Freedom	P-value
Second grade (n=932 students)			
Intercept by Lower Third	3.23	2	.20
Intercept by Upper Third	3.03	2	.22
Lower Third by Upper Third	0.61	2	>.50
Third grade (n=914 students)			
Intercept by Lower Third	3.13	2	.21
Intercept by Upper Third	2.73	2	.25
Lower Third by Upper Third	1.14	2	>.50
Fourth grade (n=770 students)			
Intercept by Lower Third	0.34	2	>.50
Intercept by Upper Third	0.54	2	>.50
Lower Third by Upper Third	0.30	2	>.50
Fifth grade (n=740 students)			
Intercept by Lower Third	1.80	2	>.50
Intercept by Upper Third	4.46	2	.11
Lower Third by Upper Third	3.14	2	.21

*** p <.001, ** p <.01, * p <.05

Limitations

I have interpreted the findings with caution as there are several limitations. First, there were methodological challenges. I ran into methodological challenges during multiple imputation because of the large amount of missing data. Imputation took place at the beginning of the analysis and choices were made that may have impacted the results. For example, for highly correlated variables, I either combined those using principal components analysis or dropped all but one that remained to represent the groups of variables. Furthermore, I only used the first principal component which represents only part of the variation in the original group of variables. It is unclear how and to what degree these choices affect the results reported here.

I have mentioned a second methodological challenge throughout this chapter related to test forms. Data used in this study were collected as part of SII's research. Their procedures for measuring student achievement were such that students who were expected to outperform or underperform the assigned test level were given an alternate level which contained different items. For SII, students' scores were comparable because the test publisher had access to linking items and item information and provided SII with the scores. This study was a secondary investigation, and, by the time I had defined this research, data had already been collected. My interest was in only multi-digit

computation items and students' scores on those items, unfortunately, the test publisher did not provide scores on only those items, linking items, or a linking group(s) of students. Given the lack of methods for handling this limitation, I was left with focusing my analysis on only students' scores that are comparable. Consequently, I had to limit my analysis to only those scores from students who took the standard test level. There are no records on what SII knew about the students' achievement for those who took the alternate form. Some students may have been assigned for an alternate test level in error, but I expect that the majority of these students were high achieving or low achieving. Therefore, this study encompassed more average students.

Another limitation relates to the TerraNova items. While the items are not reported here, very few items on each level measured only computational skills. Most items required knowledge in more than one mathematics content area. For example, typical items required students to read a graph, compute, and then select the answer. It is possible that students could compute correctly but did not read the graph correctly, and then, in the end, chose the wrong answer. Therefore, results on wrong answer are inconclusive. That is, it is not clear how effective the instructional approach was when a student gets these kinds of item incorrect.

Another limitation is related to defining classes as receiving common and blended instructional approach. I took into account how

often, day to day, students worked on multi-digit computation procedures and concepts. A piece of information that is possibly equally important in defining "dose" is how much in a day a class worked on this content. The Daily Mathematics Teacher Log asked about daily time but teachers recorded the total time on any mathematics topic. Since on many days, students worked on multiple topics, time on multi-digit computation was not discernible and not accounted for when I defined classed on instructional approach. Therefore, classes may have been unfairly assigned as receiving an approach. Without having information about daily time, it is hard to know how it might have changed a class' assignment to an approach. I leave this to future studies.

Lastly, the causal results are limited by the usual limitations of causal methodologies. During propensity score analysis, I attempted to include all potential confounding variables. The results, however, could be compromised if an unmeasured confounder was omitted.

In Chapter V, I discuss this study's implications, conclusions and future research.

Chapter V

Discussion

The purpose of this dissertation was to address a critical instructional problem by investigating methods that may well be useful for establishing scientific answers to questions about its use and effectiveness. On the surface, I have focused this investigation on comparing two instructional approaches to teaching multi-digit computation procedures and concepts in second through fifth grades. I first studied relationships between key class, teacher, and school characteristics to the use of these two approaches and how the relationships compare across grades. Then, using a causal methodological approach, I have studied whether these two approaches differ in their effects on student achievement for different groups of students. Inside this substantive study on the two approaches to teaching multi-digit computation, I examine measurement and methodological challenges to using daily teacher logs to measure treatment assignment and define instructional approach and challenges to using general mathematics test items to measure knowledge on multi-digit computation, a specific content area. In this chapter, after first

summarizing my findings, I discuss the empirical results of the substantive investigations in light of policy implications and appraise the success and challenges to measures and methods used. In each of these discussions I point out areas in need of future research.

Overall, I found that items from the SII daily teacher log are useful in measuring and distinguishing instructional approaches differing in emphasis on procedural and conceptual knowledge. From descriptive statistics, I found that teachers do differ in their emphasis on the two approaches under study and the differences align with past research findings focused on mathematics instructional approaches in U.S. elementary classrooms. Additionally, very few class, teacher, or school characteristics are predictive of the instructional approach used. This finding is also confirmed by previous studies looking at similar relationships in mathematics education. On the few significant factors, I found that the influential factors suggest that policies may be needed that focus on grades that differ from our present procedures. Then, in terms of findings from the causal investigation, I found no significant treatment effect for the blended instructional approach and no significant differences in treatment effect for student groups differing in prior knowledge. Furthermore, I found no evidence that one approach – common or blended – is superior and no evidence that supports choosing one of these instructional approaches based on students' prior knowledge. Therefore, this investigation found no support for the

instructional approaches supported by the *Standards*. I discuss each of these findings in detail below.

The Case of Common versus Blended Instructional Approaches

My findings complement the body of research and knowledge on the teaching and learning of multi-digit computation by describing general patterns on factors influencing instructional approach and by providing prescriptive information or "a recipe." Before discussing my findings related to the research questions, I have some incidental findings that I would like to comment on. In particular, the descriptive results on the teacher log items measuring instruction on procedures and concepts suggest that rather little emphasis is placed on procedures and concepts during instruction on multi-digit computation. On average, second through fifth graders spend between 26 percent and 30 percent of lessons on multi-digit computation on procedures and between 17 percent and 24 percent of these lessons on multi-digit concepts. See Table III.5. This seems like a rather small percent of lessons on content that is fundamental to students' success in Algebra. Furthermore, from previous reports from the SII study, instruction on operation, in general, occurred on 40.0 percent and 41.9 of lessons for third and fourth grades, respectively (Rowan, et al., 2004). Therefore, in classrooms from schools that have a 180 day school year, one might expect the average class of third graders, for example, to spend around only 18 days per school year

working on procedural skills and about 12 days working on conceptual skills. These days may in fact overlap leaving somewhere between 150 to 162 “allocated” days for instruction, for the average third grade class, when additional work on this content could occur. We know that instruction does not occur on of the allocated days and that, on average, no instruction on mathematics occurs on 43 days of the school year due to teacher or student absences, field trips, testing, etc. Therefore, in this average third grade class, there are between 107 to 119 days when instruction should be possible (Phelps, Corey, DeMonte, Harrison, & Ball, 2012).

These basic descriptive findings suggest that instructional guidance is needed for both day to day instructional activities and across the school year. I found a large variability in endorsing procedures and concepts, suggesting that some teachers are highly endorsing this content but these same teachers ignore this content on too many school days.

Keep in mind that this is the context in which I identified treatment classes and tested for treatment effects. Still, there are findings to report.

Discussion of Predictive Model Results

Overall, the data I present show a pattern for factors influencing instructional approach that align with middle and upper elementary

grades. The results from the unconditional models that probe grade level probabilities in receiving blended or common instructional approach suggest that fourth and fifth grade teachers find it easier to endorse the blended instructional approach that second and third grade teachers. Results from the conditional models suggest that these four grades differ in their influences but the middle grades (second and third grades) and upper grades (fourth and fifth grades) share factors of influences. On average, the instructional approach used in middle grade classes is influenced by factors outside the classroom, but in upper grades instructional approach is influenced by class characteristics, particularly student and teacher gender and ethnicity. At the basic level, these findings debunk the use of one set of policies or instructional guidance for all elementary grades. Given the different influential factors, there is reason to think about creating policies or guides that "push on" these factors differently depending on the grade(s) targeted by the policies or guides. In particular, policies and instructional guidance often lump third grade with fourth and fifth grades under a general category of upper elementary grades. These findings suggest caution in doing so, since policies on instruction may adversely affect instructional approach and thereby students' opportunities to learn. Still, in general, other results suggest that policies on instructional approach could be beneficial, if focused on the right grades. The results suggest that beginning in fourth grade, biases begin to influence instructional

approach. Policies that target instructional approach in fourth grade could eliminate the effects of bias.

Alternatively, these results suggest that middle elementary teachers need more or use more instructional guidance than the teachers of upper elementary grades. Guidance seems invaluable in this case, as it is already in use. It just needs to be refined in ways the move instruction so it aligns with reform expectations. In particular, consider the results on the effect on instructional approach of length of school days. I found a negative significant effect suggesting that classrooms in schools with long days are less likely to receive the blended instructional approach. Therefore, teachers adapt to long school days by "dumming down" their instructional approach even when good collegial support is available and when quality curricular materials are in use. Therefore, on one hand, caution is advised for schools considering lengthening their school day, but, on the other hand, policies guiding the use of the school day could be invaluable to limiting the use of such adaptations. In all, these results support the need for more instructional guidance.

Regarding the general finding that very few variables (factors) explain instructional approach when focused on multi-digit computation. This finding, the small number of significant variables, is similar to what others studying instruction have found. But, the sensibility of my findings (not the number of significant variables)—the significant effects by grade, the direction of their effects and the non-significant factors by

grade—speaks to the appropriate approach to studying instruction. It seems reasonable that future research would focus on instruction within a specified content and that lumping all content together has proved to be less informative.

Discussion of Causal Model Results

Overall, at the level of at least one day per week (or one-fifth of allocated school days) on procedures and concepts, the blended instructional approach produces no significant difference in student achievement, as compared to the common approach. This finding is consistent with previous research on student achievement suggesting that student achievement lags behind expected levels even in the post reform era. The basic descriptives discussed above—the low percentage of days when multi-digit computation is the focus of instruction and then the low percentage of days when procedures or concepts are the focus—suggests this finding is reasonable because emphasis is too low to benefit any group of students' learning, regardless of which instructional approach is used. Therefore, the treatment, blended instruction, as it is measured in this study provides too little opportunity to learn to assess its benefit over the common instructional approach, regardless of the student group membership. Furthermore, in terms of identifying the frequencies in which students should be working on multi-digit procedures and concepts, in order to develop the rich computational

knowledge envisioned in the *Standards* and noted in the comparative studies, these findings suggest that instruction should likely encourage student work on more than one day per week. This needs to be tested, but we need sufficient numbers of classes deploying a strong emphasis in order to test its effectiveness. Furthermore, given the low number of classes where emphasis occurs on more than one day per week, it may not be possible to test greater and greater emphasis using this sample which leads to questions about if and whether such approaches are in use in high poverty U.S. elementary schools, or in any U.S. elementary school. This needs to be investigated. In future research I plan to explore the use of SII data in investigating the use of treatments defined by a stronger emphasis.

Aside from studying effects on achievement on "average" students, I also set out to study the effectiveness of blended instructional approach on the academically advanced and the academically under-achieving students and whether advanced, average, or under-achieving students benefit equally from the blended instructional approach. I found no significant differences. Furthermore, I aimed to appraise if grade level differences existed. In all cases, no differences were found. These findings may in part be due to the sample characteristics. SII sampled from high poverty schools, and high poverty schools tend to have few true academically advanced students. Therefore, it is not clear just how academically advanced the upper third of students is on prior

achievement. Therefore, the lack of difference in instructional effects may in fact be due to all students being closer to average in prior knowledge. Future research is needed to study the effect of instructional approach for student groups defined by their prior knowledge.

Lastly, patterns of variation in instructional approach are consistent with previous research findings. Even with moderately low reliability, the instructional approaches varied less across schools than among teachers within the same school. That is, students in the same school, same grade, and studying the same content experience widely different instructional programs. And the findings from this investigation seem to suggest that this variability is not helping students. This variability in emphasis too often falls at the low end of the emphasis spectrum. There is too little emphasis on core content. Others have described the U.S. elementary curriculum as slowly paced and redundant. It, therefore, seems possible to change the face of our elementary curriculum by reprioritizing teachers' emphasis on core content like multi-digit computation.

Comments on Measurement and Methods

Measurement Using a Daily Teacher Log

I set out in this investigation to identify class level use of the blended and common instructional approaches. Research gives vivid details of both instructional approaches but recipes for their use across

an academic year are missing. In an era when so many talk about teacher accountability, I find the level of detail in instructional guidance on the teaching of multi-digit computation weak and timid. (I question how anyone can identify accountable instruction with the level of detail teachers are given about effective instructional approaches.) Therefore, in order to use the log items and define instructional approach in use, I had to interpret from the available guidance what is means to highly emphasis multi-digit procedures and concepts. To this end, I chose one day per five school days (or one-fifth of allocated instructional days) of focus on multi-digit computation as the distinction between low instructional emphasis and high instructional emphasis and I found no distinguishable differences in their effects on student achievement.

Still, with focused teacher log items, I found that the methods for defining and identifying instructional approach to be very systemic. Four binary response items proved to be the most useful. Binary items average across days sensibly and with easy in interpretation. Additionally, four items lead nicely to a two-by-two crosstabs table which is reasonable to synthesize. Using more items to measure and distinguish instructional approach is not impossible, but more clarity on blended and common instructional approaches is needed for their use. With tables larger than two-by-two, it is less clear which cells in the table represent which approach. For example, how should ties in emphasis be handled? Or when using more items, how should high emphasis in a subset of items

be interpreted when two items give mixed information? Overall, I found that using more items is more complex and to do so, we need a better understanding of the instructional approaches. Furthermore, given the complexity, I chose to focus on the use of four log items and, in future research, I plan to investigate the use of more log items.

The use if items that cut across more than one gateway section is worthy of comment. This investigation focused on four items from the Operation section, one of three gateway sections in SII's daily mathematics teacher log. The first gateway section focused on Number Concepts. Two items from this section seem relevant to instructional emphasis on multi-digit computation concepts – (1) composing and decomposing (grouping) whole numbers or decimals into tenths, ones, tens, hundreds, etc and (2) identifying the values of the places in whole numbers or decimals. Because teachers only entered a gateway section if they marked their focus was major or minor. If they marked "touched on briefly" or "not taught today," they were asked to skip the section. It was not clear whether using these two Number Concepts items would be helpful in identifying class emphasis on concepts. I expect that teachers might mark these items when they also marked working on the four items I did focus on in the Operations section. Adding information from items outside the Operations section proved complex, part due to the added number of items and also part due to the way teachers were asked to respond to items in gateway sections. Therefore, I also left

investigating the use of items from the Number Concepts section for future research.

The log items used in this study have low reliability, and, therefore, labels on instructional approach used in classrooms derived from the log items are subject to speculation. How can I confirm the instructional approach label given to SII classes using the four log items? The SII data might have a sample that could serve as a validation group. There are 407 classrooms that reported low work on both procedures and concepts. This sample and the blended and common sample were given in Table III.6. On the surface, it seems that in these classes instruction places little or no emphasis on multi-digit computation, and therefore, student achievement for students in the classrooms should reflect this lack of opportunity to learn this content. Studies comparing the common and blended samples to the 407 sample could document whether the common and blended approaches actually created opportunity to learn and whether the four log items are signally of an approach in use.

The idea that teachers deploy instruction within subtopics, like multi-digit computation, using a set approach is new territory. We do not know if teachers think about their work in this way or if they focus and choose their approach from lesson to lesson or day to day within a subtopic.

Measurement of Student Achievement

I created student achievement measures of multi-digit computation knowledge from selected items from the TerraNova. Very few items focused solely on multi-digit computation and when they do they are the typical straight computation problems. A large portion of items required students to read and use information from graphs and diagrams, and no items directly assessed students' multi-digit conceptual knowledge. How to interpret wrong answers and their relation to multi-digit computation knowledge is not clear. Therefore, the lack of treatment effect may in part be due to the pool of items used to measure student achievement. Future research could benefit from achievement measures created from items more focused on knowledge of multi-digit computation procedures and concepts.

My investigation faced a second limitation related to creating achievement measures on multi-digit computation. I initially set out to study the treatment effect on growth of student achievement. The lack of items parameters which created the link between test levels prohibited this inquiry. It may be that real differences between the two instructional approaches are not realized in one academic year. For some students, the real effects on their learning come only after two or three of receiving the same instructional approach. Only growth models will permit exploring the effects on achievement of instructional approaches as they

are experienced over many academic years which I plan to pursue in future research.

Analytic Methods

The causal models control for prior knowledge using a fall measure of achievement. Fall assessments occurred in late September through early November. For all classes, instruction was well under way. While I found no relationship between test level (or movement across levels), some of the treatment effect may have been absorbed by the Fall measure, especially classes who were assessed in November. How much of the treatment effect that can be absorbed is not well understood. For example, students may make large gains in achievement early in an approaches use or gains may be achieved more evenly or steadily across the year. Without an alternative achievement measure, one that occur prior to students exposure to an approach, it's unclear how approaches effect achievement.

SII did measure achievement every year in the Fall and Spring. Therefore, an alternative to the Fall measure is using the prior years' Spring achievement measure. This too has its drawbacks, particularly related to the SII study design. On set of students was followed from kindergarten through second grade and a second set was followed from third through fifth grades. Therefore, on drawback is the lack of prior years' achievement scores for all third graders. This is particularly

troublesome since multi-digit computation content is prominent in third grade. My investigation can only study second, fourth, and fifth grades, which could be a very significant contribution. Second, higher rates of missing data occur when trying to track students across grades. How to handle missing data becomes even more uncertain since test level is also unknown. Future research could try models using Spring prior achievement.

My last comment pertains to logistic regression and models or theories of instruction. First, I used a logistic regression causal model in this analysis which lead to comparing two instructional approaches, when, in fact, the number of approaches is unknown. Little is known about differences in instructional approach—when a new label and new model specifications are needed. A causal model using continuous or categorical variables as treatment maybe more appropriate for studies on instruction. These should be explored as advances in causal techniques permit their use.

Final Remarks

This research investigated the relationship between instructional approach and student achievement inside one year of instruction on one topic, multi-digit computation. It is likely that this relationship is far more complex, extending across content and topics and beyond a single year. One suggestion, raised at my oral defense, is to control for the instructional approach a student received in prior year(s). In the context

of the analytic models used in this research, one might control for prior knowledge, prior instructional approach, and their interactions terms. This model will still likely oversimplify the expected complex relationship between instructional approach and learning across a child's academic career. Using propensity scores methods and Rubin's Causal Framework seem promising for yielding the scientific evidence that is badly needed. These methods and the multi-level models used in this dissertation, however, do not account for *how* instruction is administered — *when* a treatment began and with what *regularity*—across academic years which maybe keys pieces to the success of an instruction treatment. Therefore, future research that seeks to better understand how an instructional approach acts on learning within a context of a child's academic career will need new or advanced statistical techniques.

Using a three-level model and studying treatment effects at the classroom level is not sufficient. Next steps in modeling will likely incorporate cross classified and survival analytic methods. I expect that cross classified methods will support models of different learning groups, classes, and environments and students' experience across many years, and by adding in survival analysis techniques we can bring into the model information about when topic coverage began and for how long the focus was maintained. This model is only a vision, a vision for identifying successful treatments and much needed instructional guidance. Future research is needed that investigates the use and limitations of these

advanced techniques for comparing instructional approaches while still yielding scientific evidence.

Footnotes

[i] Research focused on instructional approaches that are effective for students with learning disabilities has found the direct instruction approach (also referred to as explicit instructional approach) to be effective. This approach is defined on repeatedly practicing skills at a pace determined by the teacher's understanding of student needs and progress (Swanson, 2001). Direct instruction has been found to be especially successful when a child has problems with a specific or isolated skill (Kroesbergen & Van Luit, 2003). Also see (Steedly, et al., 2008).

[ii] Recently Star (2005) has questioned these definitions, principally the definition of procedural knowledge. He noted, for one, that the widely used definitions given by Hiebert and Lefevre suffer from an entanglement of knowledge type and knowledge quality and that deep procedural knowledge, "knowledge of procedures that is associated with comprehension, flexibility, and critical judgment…[which] is distinct from (but possibly related to) knowledge of concept", fits poorly into either of the definitions for procedural knowledge and conceptual understanding (Star, 2005, p. 408). Star also reasons that Hiebert and Lefevre's definition of procedural knowledge was intended for learning algorithms, not heuristic procedures, rules of thumb or more abstract procedures. While Star's argument has merit, Hiebert and Lefevre's definitions are particularly appropriate for this research since the definitions are centered on algorithms and the focus of the dissertation is on multi-digit computation algorithms. I do think that Star's distinction between the 1986 definition and deep procedural knowledge would come into play if I was assessing students' learning of procedural and conceptual knowledge and how the knowledge is used. As discussed in Chapter III, the items on the assessment used in this research do not permit distinguishing between type of knowledge learned and how it is used to answer the items, only whether students were working on multi-digit computation. Therefore, I think the distinctions made by Hiebert and Lefevre in their definitions are suitable for this research and I leave for future research the effort to distinguish between instructional approaches that emphasize procedural knowledge (as defined by Hiebert and Lefevre in 1986) and deep procedural knowledge.

[iii] See Moyer, Moyer, Sowder, & Threadgill-Sowder (1984) for example word problems and some comments on instructional practices and student difficulties with word problems.

[iv] Of the 25 participating countries, only 17 met the comparison guidelines at fourth grade.

[v] Of the 25 participating countries, only 16 met the comparison guidelines at third grade.

[vi] At this time, I have also excluded research on the teaching and learning of whole number multi-digit computation by adults and similar research on pre-service teachers (e.g. (Tirosh, 2000; Tirosh & Graeber, 1989)).

[vii] I identified the majority of literature included in this review through advice from researchers in this area. Additional materials were identified through Internet searches of the websites of these researchers and their project websites, through publications from the National Council of Teachers of Mathematics (e.g. yearbooks), and from reference lists provided by relevant articles. In general, I rejected publications whose subjects were not practicing teachers or school aged children of the U.S. Exceptions were allowed when reading the publication was advised by a prominent researcher in this field.

[viii] A cardinal number (such as 1, 2, or 3) is used in counting to indicate quantity but not order.

[ix] Note that children's two-digit number concepts seem to sequence through two-digit number representation similar to the sequence of computation strategies. First the sequence in a unitary concept of two-digit numbers where children generally count two-digit numbers by ones, then onto a sequence conception in which they count by tens and then by ones, and then onto a separate tens and ones conception in which the units of tens and the units of ones are counted separately. Finally an integrated conception arises when children construct both the sequence-ten and separate-tens conceptions and relate then to each other in an integrated sequence-separate conception (Fuson & Smith, 1997).

[x] A second item asks, "Of the mathematics time recorded in Question 1, how much time were you either the teacher or an observer of the teaching?" Correlation of time reported on the canonical log with this item is .46.

[xi] CTB McGraw-Hill also provided scale scores for Measurement, Geometry and Spatial Sense, Data, Statistics, and Probability, Patterns, Fractions, and Algebra, Problem Solving and Reasoning, and Communications. Scales scores are only provided on test levels that have 4 or more scale items. The sub-scales scores are derived from an algebraic formula, not from IRT. The algebraic formula is not publicly available.

[xii] Appendix A gives a table showing the collected data by fall and spring testing sessions and first through fifth grades.

[xiii] CTB McGraw-Hill was willing to share the item IRT properties but their lawyers and U of M's lawyers could not agree on the terms.

[xiv] Multiple imputation of missing scores is possible. However, since students could take a number of different forms, imputation process would need to generate possible levels the student could take and then possible scores on those levels. Since my use for the achievement data was not part of the original plan, I left exploring the use of imputed levels and scores for future studies.

[xv] A copy of the mathematics teacher log is available at www.sii.soe.umich.edu.

[xvi] Note that the average number of school days per year for schools in the SII study was reportedly 180 days, and for this analysis teachers who provided less than 18 logs were dropped as they provided instructional data on less than 10 percent of the year.

[xvii] Due to copyright laws, the items referenced by this study have not been reproduced in this report. Those interested in the forms and items should contact CTB McGraw-Hill.

[xviii] Note that Level 11 was also labeled "CTBS Basic Battery" and Levels 12-16 were labeled "CTBS Survey." The significance of these labels is not well described by in the Technical Bulletin but maybe related to the large difference in number of items on the tests.

[xix] Curriculum materials used in each classroom each year was not collected.

[xx] $E(Y_{ij} \mid \mu_{ij}) = \mu_{ij}$ and $Var(Y_{ij} \mid \mu_{ij}) = \mu_{ij}(1-\mu_{ij})$.

[xxi] $E(Y_{ij} \mid \mu_{ij}) = \mu_{ij}$ and $Var(Y_{ij} \mid \mu_{ij}) = \mu_{ij}(1-\mu_{ij})$.

[xxii] In comparing results between models with and without grand-mean center on these variables, I found no significant difference in estimates for coefficients and variances.

[xxiii] Depending on the results of this study, there may be reason to question this assumption of homogeneity and the longstanding characteristics of these approaches to instruction.

[xxiv] For the second grade model, I compared results from the chi-square tests for fixed effects versus robust standard errors and found differences in significance leading to different conclusions. In the case of robust standard errors, the chi-square 11.21 (degrees of freedom=4) and p-value is .02. I compared the standard errors and found difference for Strata 5 and strata 5*treatment which suggested that the model was mis-specified. I made changes to the model, including running a model with random slopes for strata 5 and strata 5*treatment, and found differences in chi-square test results were eliminated and I could then conclude no difference in strata*treatment.

Appendix A

Table of Frequencies for Students Taking the TerraNova

by Test Session, Grade, and Test Form

TerraNova Form	GRADE First	Second	Third	Fourth	Fifth	Sixth	Total
Fall							
Level 10	2082	35	0	1	0	0	2118
Level 11	130	1773	7	6	1	0	1917
Level 12	2	217	2630	149	2	0	3000
Level 13	0	20	70	1791	227	1	2109
Level 14	0	0	3	474	1498	0	1975
Level 15	0	0	0	14	379	0	393
Level 16	0	0	0	0	14	0	14
Total	2214	2045	2710	2435	2121	1	11526
Spring							
Level 10	62	1	1	0	0		64
Level 11	2138	73	20	1	0		2232
Level 12	13	1938	163	20	1		2135
Level 13	1	42	2210	229	26		2508
Level 14	0	2	349	2094	283		2728
Level 15	0	0	0	68	1714		1782
Level 16	0	0	0	0	46		46
Total	2254	2103	3372	2623	2227		11495

Appendix B

Table of Descriptive statistics for multiply imputed data

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Classroom Level Variables					
Class mean SES					
Complete	1465	.00000	.352306	-1.412	2.409
Imputed Data 1	40	-.14670	.423885	-1.208	0.718
Imputed Data 2	40	-.10264	.349885	-0.813	0.445
Imputed Data 3	40	-.14103	.367043	-0.813	1.087
Imputed Data 4	40	-.10893	.392434	-1.030	0.934
Imputed Data 5	40	.00981	.282303	-0.634	0.520
Combined results					
Imputed Data 1	1505	-.00390	.355014	-1.412	2.409
Imputed Data 2	1505	-.00273	.352513	-1.412	2.409
Imputed Data 3	1505	-.00375	.353309	-1.412	2.409
Imputed Data 4	1505	-.00290	.353722	-1.412	2.409
Imputed Data 5	1505	.00026	.350553	-1.412	2.409
Class proportion of male students					
Complete	1489	.00000	.221449	-0.674	0.630
Imputed Data 1	16	-.03562	.159186	-0.284	0.401
Imputed Data 2	16	.04621	.247380	-0.272	0.502
Imputed Data 3	16	-.08964	.218249	-0.522	0.266
Imputed Data 4	16	-.02701	.226448	-0.402	0.484
Imputed Data 5	16	.09064	.270990	-0.451	0.454
Combined results					
Imputed Data 1	1505	-.00038	.220871	-0.674	0.630
Imputed Data 2	1505	.00049	.221699	-0.674	0.630
Imputed Data 3	1505	-.00095	.221534	-0.674	0.630
Imputed Data 4	1505	-.00029	.221443	-0.674	0.630
Imputed Data 5	1505	.00096	.222119	-0.674	0.630
Class proportion of Caucasian students					
Complete	1489	.00000	.123726	-0.784	0.814
Imputed Data 1	16	-.01233	.301570	-0.784	0.595
Imputed Data 2	16	-.06449	.241194	-0.784	0.227
Imputed Data 3	16	-.01964	.262735	-0.752	0.398
Imputed Data 4	16	.00895	.322351	-0.752	0.458
Imputed Data 5	16	-.03116	.156729	-0.178	0.512
Combined results					
Imputed Data 1	1505	-.00013	.126704	-0.784	0.814
Imputed Data 2	1505	-.00069	.125576	-0.784	0.814
Imputed Data 3	1505	-.00021	.125849	-0.784	0.814
Imputed Data 4	1505	.00010	.127211	-0.784	0.814
Imputed Data 5	1505	-.00033	.124099	-0.784	0.814
Class proportion of African American students					
Complete	1489	.00000	.173396	-0.814	0.778
Imputed Data 1	16	.01576	.151338	-0.203	0.371
Imputed Data 2	16	.00113	.119169	-0.326	0.173

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Imputed Data 3	16	.02972	.145938	-0.257	0.411
Imputed Data 4	16	-.02813	.137423	-0.349	0.181
Imputed Data 5	16	.01563	.154965	-0.441	0.230
Combined results					
Imputed Data 1	1505	.00017	.173140	-0.814	0.778
Imputed Data 2	1505	.00001	.172882	-0.814	0.778
Imputed Data 3	1505	.00032	.173113	-0.814	0.778
Imputed Data 4	1505	-.00030	.173041	-0.814	0.778
Imputed Data 5	1505	.00017	.173172	-0.814	0.778
Class proportion of Hispanic students					
Complete	1489	.00000	.152456	-0.663	0.780
Imputed Data 1	16	-.00397	.191770	-0.567	0.412
Imputed Data 2	16	.03747	.165965	-0.238	0.512
Imputed Data 3	16	-.00445	.233027	-0.291	0.780
Imputed Data 4	16	-.01133	.171602	-0.329	0.347
Imputed Data 5	16	.03730	.078569	-0.024	0.257
Combined results					
Imputed Data 1	1505	-.00004	.152848	-0.663	0.780
Imputed Data 2	1505	.00040	.152594	-0.663	0.780
Imputed Data 3	1505	-.00005	.153419	-0.663	0.780
Imputed Data 4	1505	-.00012	.152613	-0.663	0.780
Imputed Data 5	1505	.00040	.151894	-0.663	0.780
Class proportion of Asian students					
Complete	1489	.00000	.075244	-0.733	0.811
Imputed Data 1	16	.02299	.063902	-0.037	0.219
Imputed Data 2	16	-.01382	.062061	-0.189	0.095
Imputed Data 3	16	-.00673	.020522	-0.079	0.010
Imputed Data 4	16	.03879	.129974	-0.024	0.508
Imputed Data 5	16	-.00569	.055319	-0.158	0.137
Combined results					
Imputed Data 1	1505	.00024	.075152	-0.733	0.811
Imputed Data 2	1505	-.00015	.075113	-0.733	0.811
Imputed Data 3	1505	-.00007	.074874	-0.733	0.811
Imputed Data 4	1505	.00041	.076065	-0.733	0.811
Imputed Data 5	1505	-.00006	.075049	-0.733	0.811
Class proportion of American Indian Students					
Complete	1489	.00000	.028591	-0.128	0.664
Imputed Data 1	16	-.00300	.012005	-0.048	0.000
Imputed Data 2	16	.00000	.000000	0.000	0.000
Imputed Data 3	16	.00950	.038017	0.000	0.152
Imputed Data 4	16	-.00390	.012298	-0.048	0.000
Imputed Data 5	16	-.00304	.012162	-0.049	0.000
Combined results					
Imputed Data 1	1505	-.00003	.028466	-0.128	0.664
Imputed Data 2	1505	.00000	.028439	-0.128	0.664
Imputed Data 3	1505	.00010	.028708	-0.128	0.664
Imputed Data 4	1505	-.00004	.028468	-0.128	0.664
Imputed Data 5	1505	-.00003	.028467	-0.128	0.664

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Class proportion of student of other ethnicity					
Complete	1489	.00000	.080889	-0.521	0.858
Imputed Data 1	16	-.02981	.082686	-0.257	0.139
Imputed Data 2	16	.04029	.079808	-0.028	0.257
Imputed Data 3	16	-.01538	.020381	-0.052	0.000
Imputed Data 4	16	-.00466	.080499	-0.257	0.117
Imputed Data 5	16	-.01226	.061511	-0.197	0.119
Combined results					
Imputed Data 1	1505	-.00032	.080938	-0.521	0.858
Imputed Data 2	1505	.00043	.080957	-0.521	0.858
Imputed Data 3	1505	-.00016	.080499	-0.521	0.858
Imputed Data 4	1505	-.00005	.080860	-0.521	0.858
Imputed Data 5	1505	-.00013	.080702	-0.521	0.858
Teacher has graduate degree					
Complete	1219	.00000	.434222	-0.947	0.778
Imputed Data 1	286	.01600	.431253	-0.889	0.750
Imputed Data 2	286	-.00944	0.44545	-0.917	0.727
Imputed Data 3	286	.03166	.412946	-0.875	0.778
Imputed Data 4	286	.02244	.424493	-0.947	0.750
Imputed Data 5	286	.05152	.419066	-0.917	0.727
Combined results					
Imputed Data 1	1505	.00304	.433562	-0.947	0.778
Imputed Data 2	1505	-.00179	.43624	-0.947	0.778
Imputed Data 3	1505	.00602	.430305	-0.947	0.778
Imputed Data 4	1505	.00426	.432340	-0.947	0.778
Imputed Data 5	1505	.00979	.431719	-0.947	0.778
Teacher major in education in undergraduate program					
Complete	1218	.00000	.396997	-0.950	0.900
Imputed Data 1	287	.00672	.396199	-0.933	0.769
Imputed Data 2	287	.02960	.40838	-0.917	0.857
Imputed Data 3	287	.03692	.409683	-0.938	0.857
Imputed Data 4	287	-.00494	.382279	-0.933	0.769
Imputed Data 5	287	.03101	.384943	-0.917	0.769
Combined results					
Imputed Data 1	1505	.00128	.396722	-0.950	0.900
Imputed Data 2	1505	.00564	.39922	-0.950	0.900
Imputed Data 3	1505	.00704	.399573	-0.950	0.900
Imputed Data 4	1505	-.00094	.394112	-0.950	0.900
Imputed Data 5	1505	.00591	.394788	-0.950	0.900
Teacher has permanent or standard certification					
Complete	1265	.00000	.350689	-0.958	0.875
Imputed Data 1	240	-.05728	.385791	-0.958	0.696
Imputed Data 2	240	.00615	.36431	-0.933	0.696
Imputed Data 3	240	-.05194	.377560	-0.923	0.667
Imputed Data 4	240	-.00892	.383738	-0.958	0.696
Imputed Data 5	240	.02929	.321534	-0.950	0.875
Combined results					
Imputed Data 1	1505	-.00913	.357000	-0.958	0.875
Imputed Data 2	1505	.00098	.35278	-0.958	0.875

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Imputed Data 3	1505	-.00828	.355489	-0.958	0.875
Imputed Data 4	1505	-.00142	.356046	-0.958	0.875
Imputed Data 5	1505	.00467	.346268	-0.958	0.875
Teacher is male					
Complete	1363	.00000	.297965	-0.500	0.966
Imputed Data 1	142	-.03708	.250445	-0.400	0.846
Imputed Data 2	142	-.00420	.29133	-0.500	0.957
Imputed Data 3	142	-.00932	.261175	-0.500	0.923
Imputed Data 4	142	.04183	.326676	-0.500	0.958
Imputed Data 5	142	.00457	.290334	-0.500	0.909
Combined results					
Imputed Data 1	1505	-.00350	.293936	-0.500	0.966
Imputed Data 2	1505	-.00040	.29725	-0.500	0.966
Imputed Data 3	1505	-.00088	.294624	-0.500	0.966
Imputed Data 4	1505	.00395	.300924	-0.500	0.966
Imputed Data 5	1505	.00043	.297162	-0.500	0.966
Teacher is Causian					
Complete	1341	.00000	.394481	-0.958	0.909
Imputed Data 1	164	.02417	.381976	-0.750	0.867
Imputed Data 2	164	.00931	.385555	-0.958	0.778
Imputed Data 3	164	-.00044	.374979	-0.722	0.909
Imputed Data 4	164	.01359	.379592	-0.909	0.889
Imputed Data 5	164	-.05211	.385357	-0.909	0.867
Combined results					
Imputed Data 1	1505	.00263	.393086	-0.958	0.909
Imputed Data 2	1505	.00101	.393403	-0.958	0.909
Imputed Data 3	1505	-.00005	.392283	-0.958	0.909
Imputed Data 4	1505	.00148	.392786	-0.958	0.909
Imputed Data 5	1505	-.00568	.393706	-0.958	0.909
Teacher is African American					
Complete	1341	.00000	.361015	-0.933	0.958
Imputed Data 1	164	-.03758	.323935	-0.933	0.833
Imputed Data 2	164	.02468	.371458	-0.778	0.923
Imputed Data 3	164	.00326	.328110	-0.600	0.933
Imputed Data 4	164	-.00556	.371710	-0.778	0.909
Imputed Data 5	164	.00642	.348657	-0.778	0.938
Combined results					
Imputed Data 1	1505	-.00410	.357253	-0.933	0.958
Imputed Data 2	1505	.00269	.362122	-0.933	0.958
Imputed Data 3	1505	.00036	.357475	-0.933	0.958
Imputed Data 4	1505	-.00060	.362073	-0.933	0.958
Imputed Data 5	1505	.00070	.359581	-0.933	0.958
Percent of students who are Limited English Proficient or English as Second Language					
Complete	1129	.00000	1.093843	-3.625	4.500
Imputed Data 1	376	.07792	1.199564	-3.417	4.462
Imputed Data 2	376	-.03740	1.086011	-3.625	4.375
Imputed Data 3	376	.02736	1.116557	-3.077	4.500
Imputed Data 4	376	-.04180	.949953	-3.417	3.696
Imputed Data 5	376	.07724	1.200259	-2.455	4.500
Combined results					
Imputed Data 1	1505	.01947	1.121290	-3.625	4.500

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Imputed Data 2	1505	-.00934	1.091652	-3.625	4.500
Imputed Data 3	1505	.00683	1.099252	-3.625	4.500
Imputed Data 4	1505	-.01044	1.059576	-3.625	4.500
Imputed Data 5	1505	.01930	1.121466	-3.625	4.500
Percent of students who are learning disabled or mentally impaired					
Complete	1117	.00000	1.009433	-2.200	4.300
Imputed Data 1	388	-.02473	1.079038	-1.857	4.250
Imputed Data 2	388	.14140	1.038946	-1.857	4.300
Imputed Data 3	388	-.10168	.991127	-2.200	4.120
Imputed Data 4	388	-.02389	1.034139	-2.200	4.120
Imputed Data 5	388	.06869	.960113	-1.833	4.120
Combined results					
Imputed Data 1	1505	-.00638	1.027522	-2.200	4.300
Imputed Data 2	1505	.03645	1.018657	-2.200	4.300
Imputed Data 3	1505	-.02621	1.005403	-2.200	4.300
Imputed Data 4	1505	-.00616	1.015568	-2.200	4.300
Imputed Data 5	1505	.01771	.997089	-2.200	4.300
Teachers' score on Content Knowledge for Teaching					
Complete	1222	.00000	.891189	-2.833	2.818
Imputed Data 1	283	.10943	.887838	-1.989	2.667
Imputed Data 2	283	.01361	.88573	-2.079	2.366
Imputed Data 3	283	.04808	.953443	-2.247	2.509
Imputed Data 4	283	-.01934	.900332	-2.833	2.136
Imputed Data 5	283	-.06391	.891846	-2.142	2.509
Combined results					
Imputed Data 1	1505	.02058	.891292	-2.833	2.818
Imputed Data 2	1505	.00256	.88989	-2.833	2.818
Imputed Data 3	1505	.00904	.903092	-2.833	2.818
Imputed Data 4	1505	-.00364	.892647	-2.833	2.818
Imputed Data 5	1505	-.01202	.891366	-2.833	2.818
First Principle component of student achievement on Mathematics, Reading, & English Language Arts					
Complete	1029	.00000	.866140	-2.982	2.407
Imputed Data 1	476	.00819	.909067	-2.729	2.407
Imputed Data 2	476	.03916	.89439	-2.808	2.407
Imputed Data 3	476	.01749	.891551	-2.531	2.407
Imputed Data 4	476	.05276	.873664	-2.982	2.407
Imputed Data 5	476	-.00954	.921986	-2.982	2.133
Combined results					
Imputed Data 1	1505	.00259	.879649	-2.982	2.407
Imputed Data 2	1505	.01238	.87507	-2.982	2.407
Imputed Data 3	1505	.00553	.873998	-2.982	2.407
Imputed Data 4	1505	.01669	.868583	-2.982	2.407
Imputed Data 5	1505	-.00302	.883888	-2.982	2.407
First Principle component of content & methods courses taken					
Complete	1146	.00000	.906098	-2.297	3.135
Imputed Data 1	359	.00210	.883942	-2.124	2.886
Imputed Data 2	359	.13697	1.04677	-2.282	3.135
Imputed Data 3	359	.03192	.903818	-2.297	2.572

Variables	Sample Size	Mean	Standard Deviation	Min.	Max.
Imputed Data 4	359	-.11464	.872339	-2.124	2.885
Imputed Data 5	359	-.00570	.913257	-2.297	2.773
Combined results					
Imputed Data 1	1505	.00050	.900571	-2.297	3.135
Imputed Data 2	1505	.03267	.94301	-2.297	3.135
Imputed Data 3	1505	.00762	.905357	-2.297	3.135
Imputed Data 4	1505	-.02735	.899203	-2.297	3.135
Imputed Data 5	1505	-.00136	.907510	-2.297	3.135
First principle component of years of teaching experience – career & at present school					
Complete	1257	.00000	.889608	-1.824	3.484
Imputed Data 1	248	-.02030	.928134	-1.574	2.968
Imputed Data 2	248	.04012	.85151	-1.583	3.158
Imputed Data 3	248	-.09703	.863086	-1.678	2.895
Imputed Data 4	248	.00375	.863757	-1.679	2.590
Imputed Data 5	248	.01896	.868126	-1.602	3.127
Combined results					
Imputed Data 1	1505	-.00335	.895787	-1.824	3.484
Imputed Data 2	1505	.00661	.88329	-1.824	3.484
Imputed Data 3	1505	-.01599	.885742	-1.824	3.484
Imputed Data 4	1505	.00062	.885118	-1.824	3.484
Imputed Data 5	1505	.00312	.885847	-1.824	3.484
First principle component of professional development in mathematics & Language Arts					
Complete	1173	.00000	.910024	-2.424	2.252
Imputed Data 1	332	-.00068	.947796	-1.943	2.157
Imputed Data 2	332	.02411	.96035	-2.424	2.189
Imputed Data 3	332	-.11771	.909560	-2.297	2.205
Imputed Data 4	332	.01624	.883944	-2.006	2.252
Imputed Data 5	332	-.00126	.949296	-2.266	2.157
Combined results					
Imputed Data 1	1505	-.00015	.918170	-2.424	2.252
Imputed Data 2	1505	.00532	.92109	-2.424	2.252
Imputed Data 3	1505	-.02597	.910929	-2.424	2.252
Imputed Data 4	1505	.00358	.904069	-2.424	2.252
Imputed Data 5	1505	-.00028	.918511	-2.424	2.252
School Level Variables[a]					
NSF Curriculum (1=yes, 0=no)					
Complete	110	.31	--	--	--
Imputed Data 1	5	.20	--	--	--
Imputed Data 2	5	.00	--	--	--
Imputed Data 3	5	.00	--	--	--
Imputed Data 4	5	.00	--	--	--
Imputed Data 5	5	.20	--	--	--
Combined results					
Imputed Data 1	115	.30	--	--	--
Imputed Data 2	115	.30	--	--	--
Imputed Data 3	115	.30	--	--	--
Imputed Data 4	115	.30	--	--	--
Imputed Data 5	115	.30	--	--	--

[a] Only variables that have imputed values are included.

Appendix C

Tables for T Statistics from Balance Tests[a] for Second through Fifth Grades on Five Imputed Data Sets

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
Imputed Data Set 1				
Class Variable				
Class SES composite (uclses)	-0.17	0.60	0.10	0.80
Proportion Male (uclmale)	1.52	-0.98	0.01	-0.12
Proportion Caucasian (uclwhit)	-0.60	-0.53	-1.51	-0.65
Proportion African American (uclblk)	-0.99	0.80	0.32	0.32
Proportion Hispanic (uclhisp)	0.61	-0.33	1.27	-0.04
Proportion Asian (uclasn)	0.86	0.03	-1.50	-0.26
Proportion American Indian (uclamrn)	-0.04	0.71	-0.67	-0.11
Proportion other ethnicity (ucloeth)	1.80†	-0.47	1.39	1.29
Graduate Degree (ugrddgr)	1.67†	1.21	1.23	0.25
Standard certificate (utq165a)	-0.05	-0.72	0.04	0.80
Percent LEP/ESL students (utq25a)	1.07	0.35	-0.07	0.35
Percent learning disabled students (utq25c)	0.84	0.95	1.33	0.63
Measure of content knowledge for teaching (uckt4eb)	1.98*	0.75	-0.39	-0.24
First principal component of average math, LA, & reading achievement (uPCfach)	0.53	0.13	0.78	0.81
First principal component course taking (uPCcrs)	2.76**	-1.33	1.31	0.64
First principal component years experience (uPCyexp)	1.64	0.31	0.46	0.34
First principal component professional development (uPCpd)	0.63	-1.93t	0.29	0.46
Undergraduate degree in education (uunded)	-0.37	0.59	-0.92	-0.67
Male teacher (umalet)	0.29	0.39	0.51	0.67
Caucasian teacher (uwhitet)	0.22	-1.12	-0.52	-0.32
African American teacher (ublktch)	0.33	1.45	1.22	1.13
School Variables				
Enrollment (enroll1)	-1.40	-0.68	0.85	0.27
Length of day (dylngt1)	-2.58*	1.31	-2.43*	-0.30
America's Choice school (Amer)	-1.43	-1.08	-0.39	0.91
Accelerated Schools (Accel)	0.30	0.54	-0.93	-0.15
Success for All school (SFA)	-0.30	0.80	2.28*	0.23

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
Comparison school (Comp)	1.61	-0.22	-0.91	-0.85
NSF math curriculum in use (nsf_cur)	1.16	-0.50	-1.11	1.14
Percent African American (prctbl1)	-0.71	0.61	-0.11	0.01
Proportion single parent families (snglpmn)	-0.38	0.62	1.06	0.49
Proportion of families without sufficient clothes (kidclmn)	-1.08	-1.36	0.79	0.58
Proportion of families that emphasize counting money (cntmny)	-1.13	-0.01	-1.24	-0.60
Average number of books in home (nobooks)	0.68	0.13	-0.60	0.76
Families reads often (oftnrd)	-0.21	-0.24	-0.28	0.17
Families emphasize percent & multiplication skills (prctmlt)	0.63	0.07	1.44	1.15
Socioeconomic status composite (sesmean)	0.43	0.37	-0.71	0.04
Teacher average content knowledge for teaching (ckt4ebm)	2.34*	-0.08	-1.23	-0.80
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	-0.61	-1.04	-2.44*	-0.16
Proportion of grade repeaters (prepeat)	-0.08	-0.20	-0.03	-0.44
Proportion of problem behaviors reported (pqbhprb)	0.32	0.27	-1.07	-1.22
Proportion of students in ESL (everESL)	1.65	-0.70	0.34	0.02
Imputed Data Set 2				
Class Variables				
Class SES composite (uclses)	-0.17	0.60	-0.03	0.73
Proportion Male (uclmale)	1.52	-0.98	0.01	-0.12
Proportion Caucasian (uclwhit)	-0.60	-0.53	-1.51	-0.65
Proportion African American (uclblk)	-0.99	0.80	0.32	0.32
Proportion Hispanic (uclhisp)	0.61	-0.33	1.27	-0.04
Proportion Asian (uclasn)	0.86	0.03	-1.50	-0.26
Proportion American Indian (uclamrn)	-0.04	0.71	-0.67	-0.11
Proportion other ethnicity (ucloeth)	1.80†	-0.47	1.39	1.29
Graduate Degree (ugrddgr)	1.66†	1.28	1.11	0.17
Standard certificate (utq165a)	0.03	-0.30	-0.04	0.54
Percent LEP/ESL students (utq25a)	1.17	0.29	-0.15	0.18
Percent learning disabled students (utq25c)	0.52	0.39	-0.89	-1.38
Measure of content knowledge for teaching (uckt4eb)	2.09*	0.94	-0.03	0.26
First principal component of average math, LA, & reading achievement (uPCfach)	0.43	0.26	0.78	0.81

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
First principal component course taking (uPCcrs)	2.20*	-1.52	2.09*	2.02*
First principal component years experience (uPCyexp)	1.67†	0.21	-0.16	-0.32
First principal component professional development (uPCpd)	0.11	-1.96†	0.64	1.05
Undergraduate degree in education (uunded)	-0.13	0.31	-1.09	-0.72
Male teacher (umalet)	0.66	0.96	0.33	0.40
Caucasian teacher (uwhitet)	0.42	-0.94	-0.49	-0.48
African American teacher (ublktch)	0.33	1.35	1.38	1.29
School Variables				
Enrollment (enroll1)	-1.40	-0.68	1.03	0.44
Length of day (dylngt1)	-2.58*	1.31	-2.30*	-0.19
America's Choice school (Amer)	-1.43	-1.08	-0.39	0.91
Accelerated Schools (Accel)	0.30	0.54	-0.93	-0.15
Success for All school (SFA)	-0.30	0.80	2.28*	0.23
Comparison school (Comp)	1.61	-0.22	-0.91	-0.85
NSF math curriculum in use (nsf_cur)	1.16	-0.50	-1.29	1.06
Percent African American (prctbl1)	-0.71	0.61	-0.18	-0.05
Proportion single parent families (snglpmn)	-0.38	0.62	1.06	0.49
Proportion of families without sufficient clothes (kidclmn)	-1.08	-1.36	0.79	0.58
Proportion of families that emphasize counting money (cntmny)	-1.13	-0.01	-1.24	-0.60
Average number of books in home (nobooks)	0.68	0.13	-0.60	0.76
Families reads often (oftnrd)	-0.21	-0.24	-0.28	0.17
Families emphasize percent & multiplication skills (prctmlt)	0.63	0.07	1.44	1.15
Socioeconomic status composite (sesmean)	0.43	0.37	-0.71	0.04
Teacher average content knowledge for teaching (ckt4ebm)	2.34*	-0.08	-1.23	-0.80
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	-0.61	-1.04	-2.44*	-0.16
Proportion of grade repeaters (prepeat)	-0.08	-0.20	-0.03	-0.44
Proportion of problem behaviors reported (pqbhprb)	0.32	0.27	-1.07	-1.22
Proportion of students in ESL (everESL)	1.65	-0.70	0.34	0.02
Imputed Data Set 3				
Class Variables				
Class SES composite (uclses)	-0.17	0.60	-0.03	0.73
Proportion Male (uclmale)	1.52	-0.98	0.01	-0.12

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
Proportion Caucasian (uclwhit)	-0.60	-0.53	-1.51	-0.65
Proportion African American (uclblk)	-0.99	0.80	0.32	0.32
Proportion Hispanic (uclhisp)	0.61	-0.33	1.27	-0.04
Proportion Asian (uclasn)	0.86	0.03	-1.50	-0.26
Proportion American Indian (uclamrn)	-0.04	0.71	-0.67	-0.11
Proportion other ethnicity (ucloeth)	1.80†	-0.47	1.39	1.29
Graduate Degree (ugrddgr)	1.54	0.77	1.18	-0.02
Standard certificate (utq165a)	0.24	-0.34	0.24	0.63
Percent LEP/ESL students (utq25a)	0.79	-0.25	0.24	0.66
Percent learning disabled students (utq25c)	1.01	0.83	0.83	0.18
Measure of content knowledge for teaching (uckt4eb)	1.59	0.56	-0.76	-0.65
First principal component of average math, LA, & reading achievement (uPCfach)	0.37	0.21	0.78	0.81
First principal component course taking (uPCcrs)	1.96†	-2.12*	1.54	1.14
First principal component years experience (uPCyexp)	1.22	-0.07	0.05	-0.12
First principal component professional development (uPCpd)	-0.05	-2.09*	1.30	1.52
Undergraduate degree in education (uunded)	-1.06	-0.32	-1.28	-1.14
Male teacher (umalet)	0.44	0.65	0.78	0.95
Caucasian teacher (uwhitet)	0.56	-0.69	0.21	0.29
African American teacher (ublktch)	0.28	1.40	0.65	0.50
School Variables				
Enrollment (enroll1)	-1.40	-0.68	0.95	0.35
Length of day (dylngt1)	-2.58*	1.31	-2.49*	-0.37
America's Choice school (Amer)	-1.43	-1.08	-0.39	0.91
Accelerated Schools (Accel)	0.30	0.54	-0.93	-0.15
Success for All school (SFA)	-0.30	0.80	2.28*	0.23
Comparison school (Comp)	1.61	-0.22	-0.91	-0.85
NSF math curriculum in use (nsf_cur)	1.16	-0.50	-1.29	1.06
Percent African American (prctbl1)	-0.71	0.61	-0.17	-0.04
Proportion single parent families (snglpmn)	-0.38	0.62	1.06	0.49
Proportion of families without sufficient clothes (kidclmn)	-1.08	-1.36	0.79	0.58
Proportion of families that emphasize counting money (cntmny)	-1.13	-0.01	-1.24	-0.60
Average number of books in home (nobooks)	0.68	0.13	-0.60	0.76

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
Families reads often (oftnrd)	-0.21	-0.24	-0.28	0.17
Families emphasize percent & multiplication skills (prctmlt)	0.63	0.07	1.44	1.15
Socioeconomic status composite (sesmean)	0.43	0.37	-0.71	0.04
Teacher average content knowledge for teaching (ckt4ebm)	2.34*	-0.08	-1.23	-0.80
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	-0.61	-1.04	-2.44*	-0.16
Proportion of grade repeaters (prepeat)	-0.08	-0.20	-0.03	-0.44
Proportion of problem behaviors reported (pqbhprb)	0.32	0.27	-1.07	-1.22
Proportion of students in ESL (everESL)	1.65	-0.70	0.34	0.02
Imputed Data Set 4				
Class Variables				
Class SES composite (uclses)	-0.17	0.60	0.18	0.83
Proportion Male (uclmale)	1.52	-0.98	0.01	-0.12
Proportion Caucasian (uclwhit)	-0.60	-0.53	-1.51	-0.65
Proportion African American (uclblk)	-0.99	0.80	0.32	0.32
Proportion Hispanic (uclhisp)	0.61	-0.33	1.27	-0.04
Proportion Asian (uclasn)	0.86	0.03	-1.50	-0.26
Proportion American Indian (uclamrn)	-0.04	0.71	-0.67	-0.11
Proportion other ethnicity (ucloeth)	1.80†	-0.47	1.39	1.29
Graduate Degree (ugrddgr)	1.33	0.64	0.83	-0.22
Standard certificate (utq165a)	0.48	-0.18	0.49	1.01
Percent LEP/ESL students (utq25a)	1.13	0.18	0.13	0.16
Percent learning disabled students (utq25c)	1.18	1.50	1.51	1.03
Measure of content knowledge for teaching (uckt4eb)	2.13*	1.20	-0.64	-0.39
First principal component of average math, LA, & reading achievement (uPCfach)	0.45	0.36	0.78	0.81
First principal component course taking (uPCcrs)	2.15*	-1.65	1.92†	1.29
First principal component years experience (uPCyexp)	1.57	0.09	0.09	-0.06
First principal component professional development (uPCpd)	0.32	-1.43	0.93	1.39
Undergraduate degree in education (uunded)	-0.07	0.73	-1.28	-0.82
Male teacher (umalet)	0.64	0.95	0.10	0.29
Caucasian teacher (uwhitet)	0.68	-0.44	0.12	0.55
African American teacher (ublktch)	-0.17	0.96	0.63	0.28

Variables	Grade 2 Uncond. t Statistic[b]	Grade 2 Cond. t Statistic	Grade 3 Uncond. t Statistic[b]	Grade 3 Cond. t Statistic
School Variables				
Enrollment (enroll1)	-1.40	-0.68	0.95	0.35
Length of day (dylngt1)	-2.58*	1.31	-2.49*	-0.37
America's Choice school (Amer)	-1.43	-1.08	-0.39	0.91
Accelerated Schools (Accel)	0.30	0.54	-0.93	-0.15
Success for All school (SFA)	-0.30	0.80	2.28*	0.23
Comparison school (Comp)	1.61	-0.22	-0.91	-0.85
NSF math curriculum in use (nsf_cur)	1.16	-0.50	-1.29	1.06
Percent African American (prctbl1)	-0.71	0.61	-0.19	-0.06
Proportion single parent families (snglpmn)	-0.38	0.62	1.06	0.49
Proportion of families without sufficient clothes (kidclmn)	-1.08	-1.36	0.79	0.58
Proportion of families that emphasize counting money (cntmny)	-1.13	-0.01	-1.24	-0.60
Average number of books in home (nobooks)	0.68	0.13	-0.60	0.76
Families reads often (oftnrd)	-0.21	-0.24	-0.28	0.17
Families emphasize percent & multiplication skills (prctmlt)	0.63	0.07	1.44	1.15
Socioeconomic status composite (sesmean)	0.43	0.37	-0.71	0.04
Teacher average content knowledge for teaching (ckt4ebm)	2.34*	-0.08	-1.23	-0.80
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	-0.61	-1.04	-2.44*	-0.16
Proportion of grade repeaters (prepeat)	-0.08	-0.20	-0.03	-0.44
Proportion of problem behaviors reported (pqbhprb)	0.32	0.27	-1.07	-1.22
Proportion of students in ESL (everESL)	1.65	-0.70	0.34	0.02
Imputed Data Set 5				
Class Variables				
Class SES composite (uclses)	-0.17	0.60	-0.02	0.73
Proportion Male (uclmale)	1.52	-0.98	0.01	-0.12
Proportion Caucasian (uclwhit)	-0.60	-0.53	-1.51	-0.65
Proportion African American (uclblk)	-0.99	0.80	0.32	0.32
Proportion Hispanic (uclhisp)	0.61	-0.33	1.27	-0.04
Proportion Asian (uclasn)	0.86	0.03	-1.50	-0.26
Proportion American Indian (uclamrn)	-0.04	0.71	-0.67	-0.11
Proportion other ethnicity (ucloeth)	1.80†	-0.47	1.39	1.29
Graduate Degree (ugrddgr)	1.77†	1.32	1.42	0.35
Standard certificate (utq165a)	-0.17	-0.83	0.21	0.80
Percent LEP/ESL students (utq25a)	1.53	0.88	0.69	0.85

Variables	Grade 2		Grade 3	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
Percent learning disabled students (utq25c)	1.09	1.30	0.84	-0.03
Measure of content knowledge for teaching (uckt4eb)	1.16	0.36	-0.59	-0.68
First principal component of average math, LA, & reading achievement (uPCfach)	0.83	0.36	0.78	0.81
First principal component course taking (uPCcrs)	2.34*	-1.39	1.06	0.39
First principal component years experience (uPCyexp)	1.66†	0.29	0.28	0.26
First principal component professional development (uPCpd)	-0.58	-2.68**	1.17	1.76t
Undergraduate degree in education (uunded)	-0.74	0.12	-1.17	-0.94
Male teacher (umalet)	0.92	1.13	0.65	0.89
Caucasian teacher (uwhitet)	0.34	-0.94	-0.28	-0.05
African American teacher (ublktch)	0.14	1.03	1.18	1.11
School Variables				
Enrollment (enroll1)	-1.40	-0.68	1.13	0.54
Length of day (dylngt1)	-2.58*	1.31	-2.43*	-0.31
America's Choice school (Amer)	-1.43	-1.08	-0.39	0.91
Accelerated Schools (Accel)	0.30	0.54	-0.93	-0.15
Success for All school (SFA)	-0.30	0.80	2.28*	0.23
Comparison school (Comp)	1.61	-0.22	-0.91	-0.85
NSF math curriculum in use (nsf_cur)	1.16	-0.50	-1.11	1.14
Percent African American (prctbl1)	-0.71	0.61	-0.25	-0.11
Proportion single parent families (snglpmn)	-0.38	0.62	1.06	0.49
Proportion of families without sufficient clothes (kidclmn)	-1.08	-1.36	0.79	0.58
Proportion of families that emphasize counting money (cntmny)	-1.13	-0.01	-1.24	-0.60
Average number of books in home (nobooks)	0.68	0.13	-0.60	0.76
Families reads often (oftnrd)	-0.21	-0.24	-0.28	0.17
Families emphasize percent & multiplication skills (prctmlt)	0.63	0.07	1.44	1.15
Socioeconomic status composite (sesmean)	0.43	0.37	-0.71	0.04
Teacher average content knowledge for teaching (ckt4ebm)	2.34*	-0.08	-1.23	-0.80
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	-0.61	-1.04	-2.44*	-0.16
Proportion of grade repeaters (prepeat)	-0.08	-0.20	-0.03	-0.44
Proportion of problem behaviors reported (pqbhprb)	0.32	0.27	-1.07	-1.22
Proportion of students in ESL (everESL)	1.65	-0.70	0.34	0.02

Variables	Grade 4 Uncond. t Statistic[b]	Grade 4 Cond. t Statistic	Grade 5 Uncond. t Statistic[b]	Grade 5 Cond. t Statistic
Imputed Data Set 1				
Class Variables				
Class SES composite (uclses)	-0.01	-0.30	0.32	0.26
Proportion Male (uclmale)	1.58	-0.79	1.30	-0.15
Proportion Caucasian (uclwhit)	-0.71	-0.61	-0.38	0.91
Proportion African American (uclblk)	-0.74	-0.55	-1.77†	0.31
Proportion Hispanic (uclhisp)	0.22	0.51	2.38*	-0.84
Proportion Asian (uclasn)	1.82†	0.94	0.65	-0.57
Proportion American Indian (uclamrn)	-1.89†	-0.52	0.15	0.96
Proportion other ethnicity (ucloeth)	0.17	0.41	0.06	-0.42
Graduate Degree (ugrddgr)	0.62	0.72	0.90	0.19
Standard certificate (utq165a)	0.28	1.09	1.29	-0.44
Percent LEP/ESL students (utq25a)	1.06	0.76	0.65	-0.63
Percent learning disabled students (utq25c)	0.21	0.38	1.46	0.86
Measure of content knowledge for teaching (uckt4eb)	-0.54	0.27	2.09*	-0.54
First principal component of average math, LA, & reading achievement (uPCfach)	-0.52	-0.01	1.01	-0.05
First principal component course taking (uPCcrs)	-0.21	0.98	1.01	0.62
First principal component years experience (uPCyexp)	1.17	1.61	0.17	-0.61
First principal component professional development (uPCpd)	-0.07	0.11	1.35	1.43
Undergraduate degree in education (uunded)	-0.91	-1.27	-0.49	-0.48
Male teacher (umalet)	1.82†	-0.27	0.54	0.23
Caucasian teacher (uwhitet)	-2.55*	1.14	0.56	0.69
African American teacher (ublktch)	0.81	-1.35	-1.37	-0.49
School Variables				
Enrollment (enroll1)	0.75	-0.04	1.13	-0.05
Length of day (dylngt1)	-0.26	-0.60	-0.26	0.03
America's Choice school (Amer)	1.12	-0.19	0.01	0.45
ccelerated Schools (Accel)	-1.88†	-1.25	0.68	0.16
Success for All school (SFA)	0.33	-0.41	-0.78	0.17
Comparison school (Comp)	0.37	1.78†	0.03	-0.73
NSF math curriculum in use (nsf_cur)	1.46	-0.11	0.62	0.52
Percent African American (prctbl1)	-0.72	-0.97	-1.03	0.34

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
Proportion single parent families (snglpmn)	-0.54	-0.01	-0.82	0.67
Proportion of families without sufficient clothes (kidclmn)	0.41	0.66	1.82†	-1.24
Proportion of families that emphasize counting money (cntmny)	0.25	-0.22	-1.75†	0.20
Average number of books in home (nobooks)	0.88	0.24	-0.58	0.71
Families reads often (oftnrd)	1.45	-0.42	-1.46	0.19
Families emphasize percent & multiplication skills (prctmlt)	1.57	1.33	-0.61	0.11
Socioeconomic status composite (sesmean)	0.35	-0.19	-0.25	0.00
Teacher average content knowledge for teaching (ckt4ebm)	-0.05	0.33	0.16	-0.25
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	0.67	0.13	0.05	0.11
Proportion of grade repeaters (prepeat)	-0.24	0.36	-1.58	0.83
Proportion of problem behaviors reported (pqbhprb)	-1.59	0.43	-0.06	0.27
Proportion of students in ESL (everESL)	0.81	1.51	1.11	-0.34
Imputed Data Set 2				
Class Variables				
Class SES composite (uclses)	0.01	-0.29	0.33	0.26
Proportion Male (uclmale)	1.58	-0.79	1.30	-0.15
Proportion Caucasian (uclwhit)	-0.71	-0.61	-0.38	0.91
Proportion African American (uclblk)	-0.74	-0.55	-1.77†	0.31
Proportion Hispanic (uclhisp)	0.22	0.51	2.38*	-0.84
Proportion Asian (uclasn)	1.82†	0.94	0.65	-0.57
Proportion American Indian (uclamrn)	-1.89†	-0.52	0.15	0.96
Proportion other ethnicity (ucloeth)	0.17	0.41	0.06	-0.42
Graduate Degree (ugrddgr)	0.33	-0.12	0.85	0.01
Standard certificate (utq165a)	-0.70	-0.03	1.02	-1.10
Percent LEP/ESL students (utq25a)	0.82	0.74	0.63	-0.60
Percent learning disabled students (utq25c)	0.92	0.63	1.71†	1.23
Measure of content knowledge for teaching (uckt4eb)	-0.64	-0.35	1.77†	-0.80
First principal component of average math, LA, & reading achievement (uPCfach)	-0.65	-0.06	1.73†	0.63
First principal component course taking (uPCcrs)	0.29	0.74	0.72	0.57
First principal component years experience (uPCyexp)	1.06	1.46	0.34	-0.33
First principal component professional development (uPCpd)	0.06	0.18	1.02	1.22

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
Undergraduate degree in education (uunded)	-0.81	-0.92	-0.08	-0.07
Male teacher (umalet)	1.56	-0.39	0.22	-0.33
Caucasian teacher (uwhitet)	-2.67**	1.07	0.60	0.99
African American teacher (ublktch)	0.92	-1.28	-1.31	-0.77
School Variables				
Enrollment (enroll1)	0.75	-0.04	1.13	-0.05
Length of day (dylngt1)	-0.26	-0.60	-0.26	0.03
America's Choice school (Amer)	1.12	-0.19	0.01	0.45
Accelerated Schools (Accel)	-1.88†	-1.25	0.68	0.16
Success for All school (SFA)	0.33	-0.41	-0.78	0.17
Comparison school (Comp)	0.37	1.78†	0.03	-0.73
NSF math curriculum in use (nsf_cur)	1.46	-0.11	0.62	0.52
Percent African American (prctbl1)	-0.72	-0.97	-1.03	0.34
Proportion single parent families (snglpmn)	-0.54	-0.01	-0.82	0.67
Proportion of families without sufficient clothes (kidclmn)	0.41	0.66	1.82†	-1.24
Proportion of families that emphasize counting money (cntmny)	0.25	-0.22	-1.75†	0.20
Average number of books in home (nobooks)	0.88	0.24	-0.58	0.71
Families reads often (oftnrd)	1.45	-0.42	-1.46	0.19
Families emphasize percent & multiplication skills (prctmlt)	1.57	1.33	-0.61	0.11
Socioeconomic status composite(sesmean)	0.35	-0.19	-0.25	0.00
Teacher average content knowledge for teaching (ckt4ebm)	-0.05	0.33	0.16	-0.25
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	0.67	0.13	0.05	0.11
Proportion of grade repeaters (prepeat)	-0.24	0.36	-1.58	0.83
Proportion of problem behaviors reported (pqbhprb)	-1.59	0.43	-0.06	0.27
Proportion of students in ESL (everESL)	0.81	1.51	1.11	-0.34
Imputed Data Set 3				
Class Variables				
Class SES composite (uclses)	-0.01	-0.30	0.33	0.26
Proportion Male (uclmale)	1.58	-0.79	1.30	-0.15
Proportion Caucasian (uclwhit)	-0.71	-0.61	-0.38	0.91
Proportion African American (uclblk)	-0.74	-0.55	-1.77†	0.31
Proportion Hispanic (uclhisp)	0.22	0.51	2.38*	-0.84
Proportion Asian (uclasn)	1.82†	0.94	0.65	-0.57
Proportion American Indian (uclamrn)	-1.89†	-0.52	0.15	0.96

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
Proportion other ethnicity (ucloeth)	0.17	0.41	0.06	-0.42
Graduate Degree (ugrddgr)	0.02	0.19	0.58	-0.46
Standard certificate (utq165a)	0.23	1.23	1.50	-0.60
Percent LEP/ESL students (utq25a)	1.14	0.85	0.66	-0.67
Percent learning disabled students (utq25c)	0.32	0.26	1.37	0.94
Measure of content knowledge for teaching (uckt4eb)	-0.87	-0.52	2.39*	0.45
First principal component of average math, LA, & reading achievement (uPCfach)	-0.62	-0.05	0.22	-0.57
First principal component course taking (uPCcrs)	0.09	1.81†	0.66	0.28
First principal component years experience (uPCyexp)	1.14	1.58	0.38	-0.46
First principal component professional development (uPCpd)	0.20	0.00	1.21	1.40
Undergraduate degree in education (uunded)	-0.92	-1.20	-0.12	-0.32
Male teacher (umalet)	1.78†	-0.31	0.66	0.27
Caucasian teacher (uwhitet)	-2.14*	1.51	0.59	0.69
African American teacher (ublktch)	0.41	-1.74†	-1.41	-0.63
School Variables				
Enrollment (enroll1)	0.75	-0.04	1.13	-0.05
Length of day (dylngt1)	-0.26	-0.60	-0.26	0.03
America's Choice school (Amer)	1.12	-0.19	0.01	0.45
Accelerated Schools (Accel)	-1.88†	-1.25	0.68	0.16
Success for All school (SFA)	0.33	-0.41	-0.78	0.17
Comparison school (Comp)	0.37	1.78†	0.03	-0.73
NSF math curriculum in use (nsf_cur)	1.46	-0.11	0.62	0.52
Percent African American (prctbl1)	-0.72	-0.97	-1.03	0.34
Proportion single parent families (snglpmn)	-0.54	-0.01	-0.82	0.67
Proportion of families without sufficient clothes (kidclmn)	0.41	0.66	1.82†	-1.24
Proportion of families that emphasize counting money (cntmny)	0.25	-0.22	-1.75†	0.20
Average number of books in home (nobooks)	0.88	0.24	-0.58	0.71
Families reads often (oftnrd)	1.45	-0.42	-1.46	0.19
Families emphasize percent & multiplication skills (prctmlt)	1.57	1.33	-0.61	0.11
Socioeconomic status composite (sesmean)	0.35	-0.19	-0.25	0.00
Teacher average content knowledge for teaching (ckt4ebm)	-0.05	0.33	0.16	-0.25

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	0.67	0.13	0.05	0.11
Proportion of grade repeaters (prepeat)	-0.24	0.36	-1.58	0.83
Proportion of problem behaviors reported (pqbhprb)	-1.59	0.43	-0.06	0.27
Proportion of students in ESL (everESL)	0.81	1.51	1.11	-0.34
Imputed Data Set 4				
Class Variables				
Class SES composite (uclses)	0.01	-0.28	0.32	0.26
Proportion Male (uclmale)	1.58	-0.79	1.30	-0.15
Proportion Caucasian (uclwhit)	-0.71	-0.61	-0.38	0.91
Proportion African American (uclblk)	-0.74	-0.55	-1.77†	0.31
Proportion Hispanic (uclhisp)	0.22	0.51	2.38*	-0.84
Proportion Asian (uclasn)	1.82†	0.94	0.65	-0.57
Proportion American Indian (uclamrn)	-1.89†	-0.52	0.15	0.96
Proportion other ethnicity (ucloeth)	0.17	0.41	0.06	-0.42
Graduate Degree (ugrddgr)	0.77	0.83	1.07	0.54
Standard certificate (utq165a)	-0.02	0.78	1.10	-0.74
Percent LEP/ESL students (utq25a)	0.98	0.95	0.67	-0.91
Percent learning disabled students (utq25c)	1.20	1.33	1.71†	0.98
Measure of content knowledge for teaching (uckt4eb)	-0.63	-0.21	2.27*	-0.22
First principal component of average math, LA, & reading achievement (uPCfach)	-0.54	-0.01	-0.70	-1.24
First principal component course taking (uPCcrs)	0.56	1.88†	0.54	0.26
First principal component years experience (uPCyexp)	1.22	1.40	-0.19	-1.02
First principal component professional development (uPCpd)	-0.75	-0.41	0.44	0.48
Undergraduate degree in education (uunded)	-0.89	-0.90	-0.40	-0.82
Male teacher (umalet)	1.66†	-0.16	0.47	0.19
Caucasian teacher (uwhitet)	-2.12*	1.56	0.65	0.86
African American teacher (ublktch)	0.45	-1.72†	-1.47	-0.63
School Variables				
Enrollment (enroll1)	0.75	-0.04	1.13	-0.05
Length of day (dylngt1)	-0.26	-0.60	-0.26	0.03
America's Choice school (Amer)	1.12	-0.19	0.01	0.45
Accelerated Schools (Accel)	-1.88†	-1.25	0.68	0.16
Success for All school (SFA)	0.33	-0.41	-0.78	0.17
Comparison school (Comp)	0.37	1.78†	0.03	-0.73

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
NSF math curriculum in use (nsf_cur)	1.46	-0.11	0.62	0.52
Percent African American (prctbl1)	-0.72	-0.97	-1.03	0.34
Proportion single parent families (snglpmn)	-0.54	-0.01	-0.82	0.67
Proportion of families without sufficient clothes (kidclmn)	0.41	0.66	1.82†	-1.24
Proportion of families that emphasize counting money (cntmny)	0.25	-0.22	-1.75†	0.20
Average number of books in home (nobooks)	0.88	0.24	-0.58	0.71
Families reads often (oftnrd)	1.45	-0.42	-1.46	0.19
Families emphasize percent & multiplication skills (prctmlt)	1.57	1.33	-0.61	0.11
Socioeconomic status composite (sesmean)	0.35	-0.19	-0.25	0.00
Teacher average content knowledge for teaching (ckt4ebm)	-0.05	0.33	0.16	-0.25
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	0.67	0.13	0.05	0.11
Proportion of grade repeaters (prepeat)	-0.24	0.36	-1.58	0.83
Proportion of problem behaviors reported (pqbhprb)	-1.59	0.43	-0.06	0.27
Proportion of students in ESL (everESL)	0.81	1.51	1.11	-0.34
Imputed Data Set 5				
Class Variables				
Class SES composite (uclses)	-0.02	-0.30	0.33	0.26
Proportion Male (uclmale)	1.58	-0.79	1.30	-0.15
Proportion Caucasian (uclwhit)	-0.71	-0.61	-0.38	0.91
Proportion African American (uclblk)	-0.74	-0.55	-1.77†	0.31
Proportion Hispanic (uclhisp)	0.22	0.51	2.38*	-0.84
Proportion Asian (uclasn)	1.82†	0.94	0.65	-0.57
Proportion American Indian (uclamrn)	-1.89†	-0.52	0.15	0.96
Proportion other ethnicity (ucloeth)	0.17	0.41	0.06	-0.42
Graduate Degree (ugrddgr)	0.37	0.55	0.70	-0.24
Standard certificate (utq165a)	0.14	1.16	1.14	-1.21
Percent LEP/ESL students (utq25a)	0.80	0.66	1.02	-0.35
Percent learning disabled students (utq25c)	0.85	1.24	1.43	0.91
Measure of content knowledge for teaching (uckt4eb)	-0.53	0.09	1.89†	-0.85
First principal component of average math, LA, & reading achievement (uPCfach)	-0.51	0.01	0.93	-0.90
First principal component course taking (uPCcrs)	-0.66	0.80	0.54	0.06

Variables	Grade 4		Grade 5	
	Uncond. t Statistic[b]	Cond. t Statistic	Uncond. t Statistic[b]	Cond. t Statistic
First principal component years experience (uPCyexp)	1.07	1.51	0.59	-0.36
First principal component professional development (uPCpd)	-0.12	0.19	1.03	0.88
Undergraduate degree in education (uunded)	-0.98	-1.54	-0.29	-0.68
Male teacher (umalet)	1.86†	-0.03	0.31	-0.13
Caucasian teacher (uwhitet)	-2.27*	1.34	0.72	0.80
African American teacher (ublktch)	0.60	-1.54	-1.47	-0.63
School Variables				
Enrollment (enroll1)	0.75	-0.04	1.13	-0.05
Length of day (dylngt1)	-0.26	-0.60	-0.26	0.03
America's Choice school (Amer)	1.12	-0.19	0.01	0.45
Accelerated Schools (Accel)	-1.88†	-1.25	0.68	0.16
Success for All school (SFA)	0.33	-0.41	-0.78	0.17
Comparison school (Comp)	0.37	1.78†	0.03	-0.73
NSF math curriculum in use (nsf_cur)	1.46	-0.11	0.62	0.52
Percent African American (prctbl1)	-0.72	-0.97	-1.03	0.34
Proportion single parent families (snglpmn)	-0.54	-0.01	-0.82	0.67
Proportion of families without sufficient clothes (kidclmn)	0.41	0.66	1.82†	-1.24
Proportion of families that emphasize counting money (cntmny)	0.25	-0.22	-1.75†	0.20
Average number of books in home (nobooks)	0.88	0.24	-0.58	0.71
Families reads often (oftnrd)	1.45	-0.42	-1.46	0.19
Families emphasize percent & multiplication skills (prctmlt)	1.57	1.33	-0.61	0.11
Socioeconomic status composite (sesmean)	0.35	-0.19	-0.25	0.00
Teacher average content knowledge for teaching (ckt4ebm)	-0.05	0.33	0.16	-0.25
Average Woodcock Johnson mathematics score from kindergarten school (wj_mat1)	0.67	0.13	0.05	0.11
Proportion of grade repeaters (prepeat)	-0.24	0.36	-1.58	0.83
Proportion of problem behaviors reported (pqbhprb)	-1.59	0.43	-0.06	0.27
Proportion of students in ESL (everESL)	0.81	1.51	1.11	-0.34

[a] Results from t-test comparing associations between pre-treatment covariates and receiving common instructional approach versus blended instructional approach from 2 Level HLM models: results for unconditional models and conditional models controlling for 5 strata

[b] t statistic reported for equal variances assumed.

[c] Binary variable with Pearson Chi-Square statistic reported.

†p < .1, *p < .05, **p < .01, ***p < .001.

Bibliography

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1996). *The NAEP 1996 technical report*. (NCES 1999452). Washington, D.C.: U.S. Department of Education Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=1999452.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods Research, 28*(3), 301-309.

Anghileri, J. (1989). An investigation of young children's understanding of multiplication. *Educational Studies in Mathematics, 20*(4), 367-385.

Anghileri, J. (2006). Scaffolding practices that enhance mathematics learning. *Journal of Mathematics Teacher Education, 9*(1), 33-52.

Ansalone, G., & Biafora, F. (2004). Elementary school teachers' perceptions and attitudes to the educational structure of tracking. [Feature Article]. *Education, 125*(2), 249-258.

Baek, J. (1998). Children's invented algorithms for multidigit multiplication problems. In L. J. M. M. J. Kenney (Ed.), *The teaching and learning of algorithms in school mathematics. 1998 yearbook*. (pp. 151-160). Reston, VA: National Council of Teachers of Mathematics.

Ball, D. L. (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis, 12*(3), 247-259.

Ball, D. L., Cohen, D. K., & Rowan, B. (2002). *Study of Instructional Improvement: Mathematics Log Glossary*. University of Michigan. Ann Arbor, MI.

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal, 105*(1), 3-10.

Baroody, A. J. (1984). Children's difficulties in subtraction: Some causes and questions. *Journal for Research in Mathematics Education, 15*(3), 203-213.

Barr, R., Dreeben, R., & with Wiratchai, N. (1983). *How schools work*. Chicago: The University of Chicago Press.

Becker, J., & Morgenthaler, L. (Eds.). (1998). *Math Trailblazers*. Dubuque, IA: Kendall/Hunt.

Bell, A., Fischbein, E., & Greer, B. (1984). Choice of operation in verbal arithmetic problems: The effects of number size, problem structure and context. *Educational Studies in Mathematics, 15*(2), 129-147.

Bell, A., Greer, B., Grimison, L., & Mangan, C. (1989). Children's performance on multiplicative word problems: Elements of a descriptive theory. *Journal for Research in Mathematics Education, 20*(5), 434-449.

Bell, A., Swan, M., & Taylor, G. (1981). Choice of operation in verbal problems with decimal numbers. *Educational Studies in Mathematics, 12*(4), 399-420.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design -- For Market Research, Political Polls, and Social and Health Questionnaires (Research Methods for the Social Sciences)* San Francisco, CA: Jossey-Bass.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching: A project of the American Educational Research Association* (4th ed., pp. 328-375). New York: MacMillan Publishing Company.

Brown, J. S., & VanLehn, K. (1982). Towards a generative theory of "bugs". In T. P. Carpenter, J. M. Moser & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 117-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brownell, W. A., & Moser, H. E. (1949). Meaningful vs. mechanical learning: A study in grade III subtraction. *University Research Studies in Education, 8*.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park: Sage Publications.

Camburn, E., & Barnes, C. A. (2004). Assessing the Validity of a Language Arts Instruction Log through Triangulation. *The Elementary School Journal, 105*(1), 49-73.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' Pedagogical Content Knowledge of Students' Problem Solving in Elementary Arithmetic. *Journal for Research in Mathematics Education, 19*(5), 385-401.

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*(4), 499-531.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3-20.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education, 15*(3), 179-202.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

Carroll, W. M. (1996). Use of invented algorithms by second graders in a reform mathematics curriculum. *The Journal of Mathematical Behavior, 15*(2), 137-150.

Carroll, W. M., & Porter, D. (1998). Alternative Algorithms for Whole-Number Operations. In L. J. M. M. J. Kenney (Ed.), *The teaching and learning of algorithms in school mathematics. 1998 yearbook.* (pp. 130-140). Reston, VA: National Council of Teachers of Mathematics.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 311-329.

Cohen, D. K., & Ball, D. L. (1990a). Policy and practice: An overview. *Educational Evaluation and Policy Analysis, 12*(3), 233-239.

Cohen, D. K., & Ball, D. L. (1990b). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12*(3), 331-338.

Cohen, D. K., & Ball, D. L. (1999). Instruction, capacity, and improvement. (CPRE Research Report No. RR-043). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education (CPRE).

Cook, T. D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

Cory, B. (Ed.). (1995). *Investigations in Number, data, & Space*. Palo Alto, CA: Dale Seymour.

CTB McGraw-Hill. (1999). Teacher's Guide to TerraNova (pp. 357). Monterey, CA: CTB McGraw-Hill.

CTB McGraw-Hill. (October 1997). Technical Bulletin 1 (pp. 275). Monterey, CA: CTB McGraw-Hill.

Darling-Hammond, L. (1990). Instrucitonal policy into practice: "The power of bottom over the top.". *Educational Evaluation and Policy Analysis, 12*(3), 339-347.

Ebby, C. B. (2005). The powers and pitfalls of algorithmic knowledge: a case study. *The Journal of Mathematical Behavior, 24*(1), 73-87.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 403-434.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 4-11, 19-21.

Fuson, K. C. (1982). An analysis of the counting-on solution procedure in addition. In T. A. Romberg, T. P. Carpenter & J. M. Moser (Eds.), *Addition and subtraction: A cognitive perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Fuson, K. C. (1984). More complexities in subtraction. *Journal for Research in Mathematics Education, 15*(3), 214-225.

Fuson, K. C. (1990a). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit addition, subtraction, and place value. *Cognition and Instruction, 7*, 343-403.

Fuson, K. C. (1990b). Issues in place-value and multidigit addition and subtraction learning and teaching. *Journal for Research in Mathematics Education, 21*(4), 273-280.

Fuson, K. C. (1992a). Research on learning and teaching addition and subtraction of whole numbers. In G. Leinhardt, R. Putnam & R. A. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 53-187). Hillsdale, NJ: Lawrence Erlbaum.

Fuson, K. C. (1992b). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematic teaching and learning: A project of the National Council of Teachers of Mathematics.* New York: MacMillan.

Fuson, K. C. (2003). Toward computational fluency in multidigit multiplication and division. *Teaching Children Mathematics, 9*(6), 300.

Fuson, K. C., & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first- and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education, 21*(3), 180-206.

Fuson, K. C., & Burghardt, B. H. (2003). Multidigit addition and subtraction methods invented in small groups and teacher support of problem solving and reflection. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 267-304). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fuson, K. C., Fraivillig, J. L., & Burghardt, B. H. (1992). Relationships children construct among English number words, multiunit base-ten blocks, and written multidigit addition. In J. Campbell (Ed.), *Advances in psychology: The nature and origins of mathematical skills* (pp. 39-112). North Holland, Elsevier Science.

Fuson, K. C., & Kwon, Y. (1992). Korean children's understanding of multidigit addtion and subtraction. *Child Development, 63*(2), 491-506.

Fuson, K. C., Richards, J., & Briars, D. J. (1982). The acquisition and elaboration of the number word sequence. In C. J. Brainerd (Ed.), *Children's logical and mathematical cognition: Progress in cognitive development research* (pp. 33-92). New York: Springer-Verlag.

Fuson, K. C., & Smith, S. T. (1997). Supporting multiple 2-digit conceptual structures and calculation methods in the classroom: Issues of conceptual supports, instructional design, and language. In M. Beishuizen, K. P. E. Gravemeijer & E. C. D. M. v. Lieshout (Eds.), *The role of contexts and models in the development of*

mathematical strategies and procedures (pp. 163-198). Utretch, The Netherlands: CD-B Press/The Freudenthal Institute.

Fuson, K. C., Smith, S. T., & Lo Cicero, A. (1997). Supporting Latino first graders' ten-structured thinking in urban classrooms. *Journal for Research in Mathematics Education, 28*(6), 738-766.

Fuson, K. C., Stigler, J. W., & Bartsch, K. (1988). Grade placement of addition and subtraction topics in China, Japan, the Soviet Union, Taiwan, and the United States. *Journal for Research in Mathematics Education, 19*(5), 449-458.

Fuson, K. C., Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Olivier, A. I., et al. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education, 28*(2), 130-162.

Fuson, K. C., & Willis, G. B. (1988). Subtracting by counting up: More evidence. *Journal for Research in Mathematics Education, 19*(5), 402-420.

Gamoran, A. (2010). Tracking and inequality: New directions for research and practice. In M. Apple, S. J. Ball & L. A. Gandin (Eds.), *The Routledge international handbook of the sociology of education* (pp. 213-228). London: Routledge.

Gamoran, A. (2011). Chapter 6: Designing Instruction and Grouping Students to Enhance the Learning of All: New Hope or False Promise? In M. T. Hallinan (Ed.), *Frontiers in Sociology of Education*. Dordrecht: SpringerLink, Springer Science+Business Media B.V.

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal, 32*(4), 687-715.

Gamoran, A., & Weinstein, M. (1998). Differentiation and opportunity in restructured schools. *American Journal of Education, 106*(3), 385-415.

Gelman, A., & Hill, J. (2006a). Chapter 25: Missing-data Imputation *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 529-543): Cambridge University Press.

Gelman, A., & Hill, J. (2006b). *Data Analysis Using Regression and Multilevel/Hierarchical Models*: Cambridge University Press.

Good, T. L., & Grouws, D. A. (1977). Teaching effects: A process-product study in fourth-grade mathematics classrooms. *Journal of Teacher Education, 28*(3), 49-54.

Greer, B. (1992). Multiplication and division as models of situations. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 276-295). New York: MacMillan.

Hatano, G. (1982). Learning to add and subtract: A Japanese perspective. In T. A. Romberg, T. P. Carpenter & J. M. Moser (Eds.), *Addition and subtraction: A cognitive perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Hiebert, J. (1992). Mathematics, congitive, and instrucitonal analysis of decimal fractions. In G. Leinhardt, R. Putman & R. A. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 283-322). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Human, P., Murray, H., et al. (1996). Problem solving as a basis for reform in curriculum and instruction: The case of mathematics. *Educational Researcher, 25*(4), 12-21.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: an introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Lawrence Erlbaum.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal, 30*, 393-425.

Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction, 14*(3), 251-283.

Hilton, M. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction, 84*, 1085-1092.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Hoppe, M., Gillmore, M., Valadez, D., Civic, D., Hartway, J., & Morrison, D. (2000). The relative costs and benefits of telephone interviews

versus self-administered diaries for daily data collection. *Evaluation Review, 24*(1), 102-116.

Horton, N. J., & Kleinman, K. R. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician, 61*(1), 79-90.

Imai, K., & van Dyk, D. A. (2004). Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association, 99*(467), 854-866.

Kamii, C., & Dominick, A. (1997). To teach or not to teach algorithms. *The Journal of Mathematical Behavior, 16*(1), 51-61.

Kamii, C., & Dominick, A. (1998). The harmful effects of algorithms in grades 1-4. In L. J. M. M. J. Kenney (Ed.), *The teaching and learning of algorithms in school mathematics. 1998 yearbook.* (pp. 130-140). Reston, VA: National Council of Teachers of Mathematics.

Kouba, V. L. (1989). Children's Solution Strategies for Equivalent Set Multiplication and Division Word Problems. *Journal for Research in Mathematics Education, 20*(2), 147-158.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special needs. *Remedial and Special Education, 24*(2), 97-114.

Lampert, M. (1986). Knowing, doing, and teaching multiplication. *Cognition and Instruction, 3*(4), 305-342.

Lampert, M. (1992). Teaching and learning long division for understanding in school. In G. Leinhardt, Putnam, R., & Hattrup, R. (Ed.), *Analysis of arithmetic for mathematics teaching* (pp. 221-282). Hillsdale, NJ: Lawrence Erlbaum.

Leigh, B., Gillmore, M., & Morrison, D. (1998). Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. *Journal of Clinical Epidemiology, 51*(2), 119-127.

Lemmens, P., Knibbe, R., & Tan, F. (1988). Weekly recall and diary estimates of alcohol consumption in a general population survey. *Journal of Studies on Alcohol, 49*, 131-135.

Lemmens, P., Tan, E., & Knibbe, R. (1992). Measuring quantity and frequency of drinking in a general population survey: A comparison of five indices. *Journal of Studies on Alcohol, 53*, 476-486.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.

Lubienski, S. T. (2002). A closer look at black-white mathematics gaps: Intersections of race and SES in NAEP achievement and instructional practices data. *The Journal of Negro Education, 71*(4), 269-287.

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

McIntosh, A. (1998). Teaching mental algorithms constructively. In L. J. M. M. J. Kenney (Ed.), *The teaching and learning of algorithms in school mathematics. 1998 yearbook* (pp. 44-48). Reston, VA: National Council of Teachers of Mathematics.

Miller, K. F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States: Cross-cultural insights into the nature and course of preschool mathematical developement. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 163-178). New York, NY: Psychology Press.

Mitchell, J. H., Hawkins, E. F., Jakwerth, P., Stancavage, F. G., & Dossey, J. A. (1999). Student work and teacher practices in mathematics (NCES 1999-453). Washington DC: US Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Morrow, L. J., & Kenney, M. J. (Eds.). (1998). *The Teaching and Learning of Algorithms in School Mathematics. 1998 Yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Moyer, J. C., Moyer, M. B., Sowder, L., & Threadgill-Sowder, J. (1984). Story problem formats: Verbal versus telegraphic. *Journal for Research in Mathematics Education, 15*(1), 64-68.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). Mathematics achievement in the primary

school years: IEA's third international mathematics and science report. Chestnut Hill, MA: Boston College.

National Advisory Committee on Mathematics Education. (1975). Overview and Analysis of School Mathmatics, Grades K-12. Washington, DC: Conference Board of the Mathematical Sciences.

National Center for Education Statistics. (2009). *The nation's report card: Mathematics 2009 (NCES 2010-451)*. (NCES 2010451). Washington, DC: U.S. Department of Education Retrieved from http://nces.ed.gov/Pubsearch/pubsinfo.asp?pubid=2010451.

National Center for Education Statistics. (2011). *The nation's report card: Mathematics 2011 (NCES 2012–458)*. (NCES 2012–458). Washington, D.C.: Retrieved from http://nces.ed.gov/nationsreportcard/pubs/main2011/2012458.asp.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards of school mathematics*. Reston, VA: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven: Yale University Press.

Oakes, J. (1992). Can tracking research inform practice? Technical, normative, and political considerations. *Educational Researcher, 21*(4), 12-21.

Oakes, J., Gamoran, A., & Page, R. N. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. W. Jackson (Ed.), *The handbook of research on curriculum* (pp. 570-608). New York: Macmillan.

Page, R. N. (1991). *Lower-track classrooms: A curricular and cultural perspective*. New York, NY: Teachers College Press.

Perie, M., Grigg, W. S., & Dion, G. S. (2005). *The nation's report card: Mathematics 2005*. (NCES 2006453). Washington, D.C.: U.S. Department of Education Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006453.

Peterson, P. L. (1979). Direct instruction: Effective for what and for whom? *Educational Leadership, 37*(1), 46-48.

Peterson, P. L. (1990a). The California Study of Elementary Mathematics. *Educational Evaluation and Policy Analysis, 12*(3), 241-245.

Peterson, P. L. (1990b). Doing more in the same amount of time: Cathy Swift. *Educational Evaluation and Policy Analysis, 12*(3), 261-280.

Phelps, G., Corey, D., DeMonte, J., Harrison, D., & Ball, D. L. (2012). How Much English Language Arts and Mathematics Instruction Do Students Receive? Investigating Variation in Instructional Time. *Educational Policy, 26*(5), 631-662. doi: 10.1177/0895904811417580

Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research, 15*(3), 213-234.

Raudenbush, S. W., Bryk, A., & Congdon, R. (2009). HLM Version 6.08 for Windows. Lincolnwood, IL Scientific Software International, Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. Thousand Oaks: Sage Publications.

Rist, R. C. (1970a). *The socialization of the ghetto child into the urban school system*. Washington Univ., St. Louis, MO. Social Science Inst. : Distributed by ERIC Clearinghouse.

Rist, R. C. (1970b). Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education. *Harvard Educational Review, 40*(3), 411-451.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175-189.

Rittle-Johnson, B., & Siegler, R. S. (1998). The relationship between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.), *The development of mathematical skills* (pp. 75-110). East Sussex, UK: Psychology Press.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346-362.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*(1), 41-55.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.), *Second handbook of research on teaching: A project of the American Educational Research Association* (pp. 122-183). Chicago: Rand McNally.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom; teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart and Winston.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal, 105*(1), 103-127.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics, 6*(1), 34-58.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association, 91*(434), 473-488.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine, 127*(8 Part 2), 757-763.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schifter, D. (1999). Reasoning about operations: Early algebraic thinking in grades k-6. In L. V. S. F. R. Curcio (Ed.), *Developing mathematical reasoning in grades k-12. 1999 yearbook.* (pp. 62-81). Reston, VA: National Council of Teachers of Mathematics.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching: A project of the American Educational Research Association* (3rd ed., pp. 50-91). New York, N.Y.: MacMillian.

Shin, Y., & Raudenbush, S. W. (2007). Just-Identified Versus Overidentified Two-Level Hierarchical Linear Models with Missing Data. *Biometrics, 63*(4), 1262-1268.

Shin, Y., & Raudenbush, S. W. (2010). A Latent Cluster Mean Approach to The Contextual Effects Model with Missing Data. *Journal of Educational and Behavioral Statistics, 35*(1), 26-53.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428-444.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229-293). Hillsdale, NJ: Lawrence Erlbaum.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36*(5), 404-411.

Steedly, K., Dragoo, K., Arefeh, S., & Luke, S. D. (2008). Effective mathematics instruction. *Evidence for Education, 3*(1). Retrieved from http://nichcy.org/research/ee/math

Stigler, J. W., & Hiebert, J. (1998). Teaching is a cultural activity, *American Educator,* pp. 1-10.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: The Free Press.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society. Series B (Methodological), 55*(2), 455-466.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Swanson, H. L. (2001). Searching for the best model for instructing students with learning disabilities. *Focus on exceptionsl Children, 34*(2), 1-15.

Tirosh, D. (2000). Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *Journal for Research in Mathematics Education, 31*(1), 5-25.

Tirosh, D., & Graeber, A. O. (1989). Preservice elementary teachers' explicit beliefs about multiplication and division. *Educational Studies in Mathematics, 20*, 79-96.

University of Chicago School Mathematics Project. (1999). *Everyday Mathematics*. DeSoto, TX: McGraw-Hill.

VanLehn, K. (1986). Arithmetic procedures are induced from examples. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Lawrence Erlbaum.

Vest, F. R. (1969). A catalog of models for the operations of addition and subtraction of whole numbers. *Educational Studies in Mathematics, 2*(1), 59-68.

Vest, F. R. (1971). A catalog of models for multiplication and division of whole numbers. *Educational Studies in Mathematics, 3*(2), 220-228.

Wallen, N. E., & Travers, R. M. W. (1963). Analysis and investigations of teaching methods. In N. L. Gage (Ed.), *Handbook of Research on Teaching* (pp. 448-505). Chicago: Rand McNally and Company.

Wayman, J. C. (2003). *Multiple Imputation for Missing Data: What Is It and How Can I Use It?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). Looking inside the Classroom: A Study of K-12 Mathematics and Science Education in the United States. 356. Retrieved from http://www.horizon-research.com/insidetheclassroom/reports/looking/complete.pdf

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399.

Wiemers, N. J. (1990). Transformation and accommodation: A case study of Joe Scott. *Educational Evaluation and Policy Analysis, 12*(3), 281-292.

Wilson, S. M. (1990). A conflict of interests: The case of Mark Black. *Educational Evaluation and Policy Analysis, 12*(3), 293-310.

Yackel, E., Cobb, P., & Wood, T. (1999). The interactive constitution of mathematical meaning in one second grade classroom: An illustrative example. *Journal of Mathematical Behavior, 17*(4), 469-488.

Yuan, Y. C. Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0). 1-3. Retrieved from http://support.sas.com/rnd/app/papers/multipleimputation.pdf

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientifitc Software International.