MATHEMATICS TEACHERS' PROFESSIONAL EXPERIENCE AND THE DEVELOPMENT OF MATHEMATICAL PROFICIENCY FOR TEACHING

by

ERIK DANIEL JACOBSON

(Under the Direction of JEREMY KILPATRICK)

ABSTRACT

The purpose of this dissertation was to investigate the role of background, experience, and interactions with colleagues and students in the development of *mathematical proficiency for teaching*, operationalized as teachers' content knowledge for teaching and beliefs about learning and mathematics. I conducted three studies in different contexts. The purpose of the first study of Texas K–12 mathematics teachers in their first 5 years of teaching was to describe how mathematical proficiency for teaching multiplicative reasoning varied across preparation, school contexts, and a wide range of grade levels.

Surprising findings from the first study led to two follow-up studies: (1) content knowledge for teaching was not positively related to length of teaching experience and (2) the length of student teaching had no significant relationship to mathematical proficiency for teaching. The second study used a longitudinal design to study change in Grades 6–8 teachers' mathematical proficiency for teaching multiplicative reasoning topics. I found that teachers' content knowledge for teaching increased over the semester, especially for teachers with less mathematical preparation, but that their self-efficacy beliefs decreased. In the third study, I

found that the quality, timing, and length of student teaching were significant predictors of content knowledge and beliefs. For all three outcomes, practicum length was found to moderate the effect of timing, with early timing of student teaching having significant positive effects for prospective teachers in programs with shorter student teaching. These results will support future work designing and investigating interventions that support teachers' on-the-job development of mathematical proficiency for teaching.

INDEX WORDS: Mathematics education; educational policy; multiplicative reasoning; mathematical proficiency for teaching; content knowledge for teaching; mathematical knowledge for teaching; productive disposition for teaching; student teaching; teaching experience; prospective teachers; inservice teachers; learning to teach; teacher certification

MATHEMATICS TEACHERS' PROFESSIONAL EXPERIENCE AND THE DEVELOPMENT OF MATHEMATICAL PROFICIENCY FOR TEACHING

by

ERIK DANIEL JACOBSON

BA, Dartmouth College, 2004

MA, University of Georgia, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

MATHEMATICS TEACHERS' PROFESSIONAL EXPERIENCE AND THE DEVELOPMENT OF MATHEMATICAL PROFICIENCY FOR TEACHING

by

ERIK DANIEL JACOBSON

Major Professor:	Jeremy Kilpatrick
Committee:	Melissa Freeman Andrew Izsák Denise Spangler

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

DEDICATION

To Brian C. Henry for teaching me the courage to aspire and the persistence to achieve.

ACKNOWLEDGEMENTS

No academic work is truly solitary, and my dissertation project and doctoral education have been supported by—indeed, have only been possible because of—the generosity of many. I am deeply grateful to Dean Grasso and the Graduate School at the University of Georgia (UGA) for supporting my time at UGA with a Presidential Fellowship. My dissertation research was also supported in part by a grant from the American Educational Research Association which receives funds for its AERA Grants Program from the National Science Foundation under Grant #DRL-0941014. I note that the dissertation reflects my opinions and does not necessarily reflect those of the granting agencies.

I owe a great deal of thanks to Patricia Wilson and Jeremy Kilpatrick, who through the Center for Proficiency in Teaching Mathematics at UGA supported travel, research projects, and training—including a pilot research project at the Elementary Mathematics Laboratory (EML) at the University of Michigan. Many thanks are due to Deborah Ball, Hyman Bass, and the others at EML for all I learned there, and in particular the orientation I developed toward practice-focused teacher education. I thank Mark Thames for all I have learned about mathematical knowledge for teaching from our work together.

Most especially, I thank my dissertation committee: Denise, you have supported me in countless ways. “Reassuring” does not nearly capture what your sanity, practicality, and open office door have been for me over the last five years. Your commitment to education at all levels has had a profound effect on how I understand what mathematics education actually *is* as an endeavor, what it takes, and what my role might be. Melissa, you have been a critical and

supportive reader and have pushed me in coursework and conversations to think deeply about epistemology and ontology and how these concerns affect the research endeavor. I do not think as I once did—or at least I am more aware of *how* I think—and I take less for granted; thank you. Andrew, our work together has enriched me by shaping my image of scholarship. You have taught me what excellence in mathematics education research looks like under the hood, to be skeptical of myself, and to be strategic. Thank you for those invaluable lessons. I count myself very fortunate to have apprenticed with you. Jeremy, you have given me the space, time, and prodding to develop my own interests and ideas. Your wide engagement with mathematics education has shown me the value of seriously and critically considering other perspectives. You have taught me the importance of details and to say clearly what I mean; I will spend my career practicing those things and paying off the debt I owe as I work with others. Your generosity and encouragement have provided more support to me over the last five years than perhaps you know; I am deeply grateful.

There are many others who have had had a role in this work. Thank you to all.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES ix

LIST OF FIGURES xiii

CHAPTER

1 INTRODUCTION 1

Research Problem and Purpose 2

Research Questions 4

Significance 8

Dissertation Overview 11

2 FRAMEWORK AND LITERATURE REVIEW 13

Multiplicative Reasoning 13

Mathematical Proficiency for Teaching 22

Teachers' Professional Experience in Schools 34

3 THE TEXAS STUDY 40

Data and Methods 42

Results 61

4 THE GEORGIA STUDY 85

Data and Methods 87

Results 98

5 THE UNITED STATES STUDY 125

Data and Methods 126

Results 137

6 DISCUSSION AND CONCLUSIONS 155

Texas Study 155

Georgia Study 159

United States Study 163

Looking Across the Studies 166

Future Research 169

REFERENCES 177

APPENDICES

A CONCEPTUAL ANALYSIS OF MULTIPLICATIVE REASONING 193

B MATHEMATICAL KNOWLEDGE FOR TEACHING AND TEACHING SELF-EFFICACY INSTRUMENT DEVELOPMENT 202

C INTERVIEW PROTOCOL FOR THE GEORGIA STUDY 221

LIST OF TABLES

Page

Table 1: Summary of the Three Dissertation Studies 5

Table 2: Content Knowledge for Teaching Instruments, Constructs, Subconstructs, and Categories Classified by Three Kinds of MKT 28

Table 3: Classification of MKT Items by Pedagogical Task, Content Topic, and Problem Type 46

Table 4: Sample Items From the TSE Beliefs and the TSE Sources Instruments 47

Table 5: Auxiliary and Independent Variables Used in Multivariate Regression Analyses..... 54

Table 6: Demographic and Certification Characteristics as a Percentage of Two TX-TEDS Study Samples and the Analytic Sample..... 56

Table 7: Age and Undergraduate Selectivity of Participants by TX-TEDS Study and Analytic Sample..... 58

Table 8: Preparedness to Teach Fractions by Certification Grade Level and Participant Subgroup 60

Table 9: Correlations Among the Factors of TSE Beliefs and TSE Sources (Raw Scores)..... 66

Table 10: Tests of Multiple Group Measurement Invariance Across Early Entry Status for the Polytomously Scored Instruments: Personal Efficacy (PE), Knowledge Efficacy (KE), Quality of Practicum (QP), and Collaborative Activity (CA) 73

Table 11: Mean and Mean Differences of Quality of Practicum (QP) Item Responses by Early Entry Status..... 73

Table 12: Tests of Multiple Group Measurement Invariance Across Certification Grade Level for the Polytomously Scored Instruments: Personal Efficacy (PE), Knowledge Efficacy (KE), Quality of Practicum (QP), and Collaborative Activity (CA) 74

Table 13: Mean and Mean Differences of Personal Efficacy (PE) Item Responses by Certification Grade Level 75

Table 14: The Mantel-Haenszel (MH) and Generalized MH Chi-Square Test for Differential Item Functioning Across Certification Groupings..... 76

Table 15: Model Fit Indices for the Measurement CFA Model and SEMs of Mathematical Proficiency for Teaching..... 78

Table 16: Variance of Mathematical Proficiency for Teaching Explained by Each Model..... 79

Table 17: Wald Tests of Hypotheses Across the Three Models Predicting Mathematical Proficiency for Teaching..... 82

Table 18: Standardized Path Coefficients for the Structural Portion of Model 3, a SEM of Mathematical Proficiency for Teaching (N = 492)..... 83

Table 19: District Participation in the Study..... 89

Table 20: Number of Participants in the Analytic Sample Who Did and Did Not Participate in Different Survey Waves..... 90

Table 21: Summary Statistics for the Explanatory Variables..... 92

Table 22: Percentage of Missing Data on Key Time-Invariant and Time-Varying Variables..... 97

Table 23: Median Variance and Variance Partition Coefficients for the Individual and School Levels in 2- and 3-Level Models of MKT, PE, and KE Across 50 Imputed Data Sets.... 99

Table 24: Models of Mathematical Knowledge for Teaching (MKT) 104

Table 25: Comparisons of MKT Models Over 50 Imputed Data Sets 106

Table 26: Comparisons of PE Models Over 50 Imputed Data Sets 110
Table 27: Models of Personal Efficacy (PE) 111
Table 28: Comparisons of Knowledge Efficacy (KE) Models Over Imputed Data Sets 113
Table 29: Models of Knowledge Efficacy (KE) 114
Table 30: Percentage of Missing Data on Key Analytic Variables for the Sample, the Analytic Sample, and Those Represented by These Samples 127
Table 31: Quality of Student Teaching Measure: Frequency of Opportunities to Connect Teaching and Learning 131
Table 32: TEDS-M Measures for Teachers' Beliefs: Mathematics-as-Inquiry and Active Learning Beliefs 133
Table 33: Covariates Used to Mitigate Selection Bias 135
Table 34: Fit Information for the Models of Teacher Outcomes 139
Table 35: Partitioned Level 1 (Individual) and Level 2 (Program) Variance in Teacher Outcomes 141
Table 36: Predictors of Teachers' Pedagogical Content Knowledge and Content Knowledge . 143
Table 37: Predictors of Teachers' Beliefs 150
Table B1: Numbers of Respondents With Omissions in the MKT Instrument Response Data . 203
Table B2: The Log-Likelihood Ratio Test and Model Fit Indices Comparing the 1PL and 2PL Models for All 26 MKT Items 204
Table B3: Increases in Item and Person Parameter Estimates From the Original to the Revised 25-Item 2PL Model 207
Table B4: Item Parameters for the Final MKT Instrument 208

Table B5: Distribution of MKT Items by Pedagogical Task, Content Topic, and Problem Type on the MKT and Short MKT Instruments 209
Table B6: Item Parameters for the Short MKT Instrument 210
Table B7: Number of Respondents With Omissions in Response Data for the TSE (Both TTMR Surveys, N=492) and TSE Sources Instruments (Second TTMR Survey, N=386) 212
Table B8: Univariate Statistics for Item Data From the TSE Beliefs Instrument 215
Table B9: Univariate Statistics for Item Data from the TSE Sources Instrument 216
Table B10: Model Fit Indices (and 95 Confidence Intervals) for the CFA Models of the Teaching Self-Efficacy (TSE) and TSE Sources Instruments and for the SEM Regressing TSE on Its Sources 217

LIST OF FIGURES

Page

Figure 1: Sample (a) dichotomous response and (b) multiple-choice items used in the MKT instrument 45

Figure 2: Histograms of the independent variables for student teaching length and teaching experience 51

Figure 3: Histograms of participant age for nonresponding volunteers and the analytic sample 59

Figure 4: The score distribution and standard error of estimate over the range of observed ability (θ) for the MKT instrument 65

Figure 5: Structural equation model for the factors of TSE Beliefs regressed on the TSE Sources factors. All bold path coefficients were statistically significant at $p < .01$ except KE on VE which was close to statistical significance at the $p < .05$ level ($p = .06$, not shown). 67

Figure 6: The common measurement model used to test invariance across certification grade level and early entry status groupings..... 72

Figure 7: Histograms showing the distribution of student teaching length in programs with normal timing (last year only) and early timing (before the last year) for student teaching; the reported frequencies are population estimates based on sampling weights..... 130

Figure 8: The effect of early timing student teaching on content knowledge with 95% confidence bands as compared to predicted knowledge after 16 weeks of normal timing student

teaching (marked ×); early timing had a statistically significant effect when student teaching was less than 16.4 weeks or more than 26.7 weeks. 148

Figure 9: The effect of early timing of student teaching on math-as-inquiry beliefs with 95% confidence bands as compared with predicted beliefs after 16 weeks of normally timed student teaching (marked ×); early timing had a statistically significant positive effect when student teaching was less than 15.8 weeks..... 152

Figure 10: The effect of early timing of student teaching on active learning beliefs with 95% confidence bands as compared with predicted beliefs after 16 weeks of normally timed student teaching (marked ×); early timing had a statistically significant positive effect when student teaching was less than 14.6 weeks..... 153

Figure A1: A double number line representation for the isomorphism of measures multiplicative structure..... 195

Figure A2: The multiplicative structure that defines the fraction *A/B* as an extensive quantity 196

Figure A3: Multiplicative structures for (a) the product $P = Q/R$ of the fractions K and $G = M/N$, (b) the fraction $P = Q/R$, and (c) the fraction $G = M/N$..... 198

Figure A4: In (a) sharing division the quotient is an equivalence class *[X]*; in (b) measurement division the quotient is a quantity of groups *Y*..... 198

Figure A5: An elaborated illustration of the isomorphism of measures multiplicative structure for the simple proportion $A/C = B/D$ showing the rate or constant quotient K, the equality of the scale factors C/A and D/B, and the associated homogenous linear functions f and its inverse f'..... 199

Figure A6: A multiplicative structure for the 3 people, 2 hours pair (a) ; a multiplicative structure for the 2 people, 3 hour pair (b); and a multiplicative structure for $K \times G = P$ when all three are extensive quantities (c)..... 201

Figure B1: Item 20 on the MKT instrument, which did not function well with the rest of the items..... 205

Figure B2: Measurement model for the 12-item Teaching Self-Efficacy instrument. All path coefficients were statistically significant at $p < .01$. Measurement error estimates are proportions of unexplained variance..... 219

Figure B3: Measurement model for the 15-item Sources of Teaching Self-Efficacy instrument. All path coefficients were statistically significant at $p < .01$. Measurement error estimates are proportions of unexplained variance..... 220

CHAPTER 1

INTRODUCTION

In the United States, Grades K–12 mathematics teachers spend anywhere from a handful of weeks up to 5 years in preparation before accepting responsibility for a critical portion of the school learning of 20 to 200 students each year. This preparation often includes *student teaching,* a practicum or field-based training experience in school classrooms during which prospective teachers take on the work of teaching in increasingly complex and authentic ways. Many teachers entering the profession through alternative routes have little or no student teaching experience; for these teachers, the bulk of the learning that might happen during traditional student teaching must happen on the job, if it happens at all. Thus, schools and teachers' professional activity are an important context for teachers' and prospective teachers' development and change.

Teachers (and prospective teachers) in schools learn from other teachers, from school and university mentors, in concurrent university coursework, at professional development workshops, and from their own classroom experiences working with students. The multitude of sites for teacher learning in schools illustrates just how complex and multifaceted the problem of understanding teachers' learning really is. There can be no single solution to this problem for the simple reason that there is no single experience.

The variation in prospective and new teachers' preparation experiences (of which the overall length of preparation is but a single feature) is the consequence of a wide range of policies at the federal, state, and district levels—policies that are based on differing assumptions

about what is necessary or propitious for teacher preparation and development. For over a decade, there have been persistent calls for more research on the effects of various features of teacher preparation programs including student teaching (e.g., Wilson, Floden, & Ferrini-Mundy, 2002; Wilson et al., 2009). Little can be said with confidence about how teachers *should* learn because little is known about how teachers *do* learn. At the same time, the wide range of teacher education experiences provides an opportunity to explore the question of how teachers learn across a variety of contexts. In this dissertation, I report empirical research that compares learning outcomes across contexts with the aim of providing evidence to support both empirically-based and theoretical arguments that are relevant to the question of how teachers should learn, the question that current polices already seek to answer.

Research Problem and Purpose

The research reported in this dissertation lies at the intersection of teacher education policy and mathematics education research. Policy differences at the state and district level contribute to the variation in context that affords the comparisons made in the three studies reported in the present dissertation, and the results of the research have implications for policy. In a similar fashion, theoretical accounts of how mathematics teachers learn provide hypotheses explored in the studies reported in the dissertation, and the results of those explorations feed back to inform and refine theoretical accounts of learning. In an effort to find evidence of the sites and processes of conceptual change that matter most across teachers' experience, the present studies provide a wide view rather than all focusing on a single process, mechanism, or occasion for teacher development (e.g., mentoring). Because of prior research on how teachers help students learn mathematics in schools, the common teacher outcome across all studies was mathematical proficiency for teaching. Kilpatrick, Swafford, and Findell (2001) argued that

mathematics teaching proficiency involves several interrelated components. I focus on two: content knowledge and "a productive disposition toward mathematics, teaching, learning, and the improvement of practice" (p. 380). *Productive disposition for teaching mathematics* can thus be defined as mathematics teachers' orientation toward—and their related beliefs and attitudes about—the subject of mathematics, teaching and learning it, and their own professional growth. This dissertation reports evidence pertaining to how teachers develop mathematical proficiency for teaching from their professional experience in schools.

I hypothesized interactions between teachers' content knowledge and beliefs, and I addressed aspects of a broad research problem: How do teachers develop mathematical proficiency for teaching in the context of professional experience? A promising, unexplored site for this research is the changing practice of student teachers and early entry teachers (who begin teaching before completing the requirements for certification) because conceptual and dispositional change are likely to be more extreme and therefore more evident with these teachers than with others.

My research program is structured loosely on recommendations for a research agenda that includes "proofs of concept, then broader studies of conditions of effectiveness, and finally experiments to determine whether interventions can produce conditions of effectiveness in various contexts" (Heck, Weiss, & Pasley, 2011, pg. 5). These recommendations came from a team at Horizon Research tasked with identifying an agenda for research in response to the national mathematics standards released in 2010 by the National Governors Association Center for Best Practices and the Council of Chief State School Officers. To generate their recommendations for research, the team solicited input from mathematics education and

educational policy researchers. In the report, Heck and colleagues (2011) argued that the three kinds of recommended studies had "different and complementary" roles:

> Proofs of concept are generally opportunistic, or conducted under fairly special circumstances; they can be used to establish the viability of a particular relationship. Broader studies of the conditions of effectiveness extend beyond those special circumstances to examine the range of conditions under which particular relationships exist, and for whom. These studies can also offer explanations of why and how particular conditions result in various outcomes. Finally, experiments/quasi-experiments are useful for establishing whether creating particular conditions in fact facilitates the relationships of interest, and for whom. (p. 5)

In a proof-of-concept study, I compared two cases of early entry teachers who participated in a professional development workshop and found evidence that the pedagogical content knowledge for fraction and proportions they learned in the workshop was mediated by their professional goals, their attention to student thinking, and their habits of collaboration with other teachers (Jacobson, in press). This dissertation includes three studies of the "conditions of effectiveness" of teaching experience on subsequent conceptual and dispositional change. The purpose of the studies was to investigate the role of background, experience, and interactions with colleagues and students in the development of mathematical proficiency for teaching. The results will support future work in designing and investigating interventions that support teachers' on-the-job development of mathematical proficiency for teaching.

Research Questions

The three studies reported in this dissertation all address aspects of the same broad problem stated above. Each study dealt with a different population of participants and used different measures and different psychometric models to operationalize mathematical proficiency for teaching (see Table 1). In this section, I describe the purpose, population, and research question for each study. Rationales for the design of each study are given at the beginning of Chapter 3, 4, and 5.

Table 1.

Summary of the Three Dissertation Studies.

	TX Study (Chapter 3)	GA Study (Chapter 4)	U.S. Study (Chapter 5)
Purpose	• Validate domain-specific measures • Describe variation in mathematical proficiency for teaching	• Explore change in mathematical proficiency for teaching	• Describe the features of student teaching associated with mathematical proficiency for teaching
Participants	• K–12 math teachers from Texas in 1st to 6th year	• Grades 6–8 mathematics teachers from Georgia	• Prospective K–6 teachers in U.S. public institutions
Outcomes			
Teachers' content knowledge	• Mathematical knowledge for teaching multiplicative reasoning topics (25 items)	• Mathematical knowledge for teaching multiplicative reasoning topics (17 items)	• Mathematics content knowledge (74 items) • Mathematics pedagogical content knowledge (32 items)
Teachers' beliefs	• Teachers' self-efficacy beliefs for teaching fraction, ratio, and proportion topics (11 items)	• Teachers' self-efficacy beliefs for teaching ratio and proportion topics (11 items)	• Teachers' beliefs about mathematics as an inquiry activity (5 items) • Teachers' beliefs about active learning in mathematics (6 items)
Psychometric models used	• Confirmatory factor analysis and structural equation models of latent traits	• Raw scores and a two parameter logistic item response theory model (a latent trait model)	• One and two parameter logistic item response theory models (latent trait models)

The purpose of the first study was to validate measures of teachers' content knowledge and beliefs that focused on the content domain of multiplicative reasoning. This focus was strategic because teachers' content knowledge of and disposition toward that domain can inform good teaching across Grades K–12, but the content is primarily taught in Grades 4 to 7. Focused instruments that are sensitive to domain-specific mathematical proficiency for teaching can support the investigation of hypotheses related to domain-specific experience that cannot be done with existing, coarser instruments. The second purpose of the first study was to describe how mathematical proficiency for teaching multiplicative reasoning varied in a sample of *practicing*

teachers who had diverse preparation experiences and who were working in a wide range of grade levels.

In the first study, I collected survey data from teachers of Grades K–12 mathematics in Texas certified in the previous 5 years ($N = 492$). The research questions for the first study were as follows:

1. How valid and reliable are the content knowledge and teaching self-efficacy measures adapted for the domain of multiplicative reasoning?
2. How does mathematical proficiency for teaching multiplicative reasoning vary (a) with academic preparation, student teaching, and teaching experience and (b) by early entry status and across certification grade level?

Surprising findings from the first study led to the second and third studies reported in this dissertation. These results were that (1) mathematical knowledge for teaching was not positively related to length of teaching experience and (2) the length of student teaching had no significant relationship to mathematical proficiency for teaching.

I addressed limitations in the design of the first study that may have affected these findings in the follow-up studies. In the first study, I compared different cohorts to describe the relationship between length of teaching experience and mathematical proficiency for teaching, a strategy susceptible to cohort bias. In the second study, I addressed this limitation by using a longitudinal design. In addition, the first study lacked adequate controls for differences among teachers that may have been common causes of program selection (and hence the length of student teaching) and the outcome measures. The data I used in the third study included covariates that have been used by other researchers in similar kinds of analyses to mitigate selection bias (e.g., Ronfeldt & Reininger, 2012).

The purpose of the second study was to study *change in mathematical proficiency for teaching* longitudinally using the domain-specific measures validated in the first study, and to investigate how that change was related to features of teachers' professional experience and especially to teachers' interactions with colleagues and students. The data I used in the second study came from three waves of surveys of 199 Grade 6–8 mathematics teachers in Georgia over one semester and interviews of 17 of the surveyed teachers. The research questions for the second study were as follows:

3. What features of teachers' experience in schools are associated with change in mathematical proficiency for teaching multiplicative reasoning topics?
4. How does change in mathematical proficiency for teaching multiplicative reasoning topics differ (a) across schools, (b) between teachers who teach that content and those who do not, and (c) with the frequency of teachers' collaboration and collegial activity focused on student learning?

The purpose of the third study was to examine how features of *student teaching* affect mathematical proficiency for teaching. In the study, I used data collected from a nationally representative sample of public universities preparing teachers in the United States (Teacher Education and Development Study in Mathematics [TEDS-M], Tatto et al., 2012), and in particular the available sample of Grades K–6 prospective teachers ($N = 1044$). The research question for the third study was as follows:

5. How are the timing, length, and quality of student teaching related to prospective teachers' mathematical proficiency for teaching the K–6 curriculum?

Significance

The significance of the research problem addressed in this dissertation follows from the mediocre ranking of United States students and teachers in international comparisons, the growing recognition of teachers' critical role in improving the quality of mathematics education, recent progress in identifying what makes teachers effective, and lackluster results of professional development aimed at improving teachers' content knowledge over the last decade. Moreover, national changes in how teachers are prepared have led an ever-increasing number of new teachers to rely on schools and districts for professional training. Little is known about how best to support these teachers' induction and professional growth.

Persistent and widespread concern for inadequate student achievement and poor international standing have been used as rallying calls for educational reform and driven decades of policy aimed to improve mathematics education in the United States. These efforts have been spurred on by bleak national assessments (e.g., National Commission on Excellence in Education, 1983; National Mathematics Advisory Panel [NMAP], 2008) and continued poor performance in international comparison exams (e.g., National Center for Education Statistics, 2008). Teachers are increasingly seen as a key lever for improving educational quality. Recent legislation and related policy initiatives in the United States have addressed concerns about mathematics teachers, including the No Child Left Behind Act of 2001, which stipulated "highly qualified" teachers in Title-I schools, and the American Recovery and Reinvestment Act of 2009, which encouraged improved teacher effectiveness through the Race to the Top funding competition.

Definitions of effective teaching vary widely, as do recommendations for necessary teacher qualifications (Goe, Bell, & Little, 2008). The lack of consensus may be traced to

decades of mixed and uneven results from research on the relationship between teacher characteristics and student achievement (Hanushek, 1986, 1996; Wayne & Youngs, 2003). In survey-based research, methodological and statistical limitations often prevent strong causal claims, and uncontrolled selection bias can precipitously affect results. Teacher educators conduct primarily qualitative research on teacher characteristics and face limitations that often preclude generalizing results beyond the cases they study (Kennedy, 2008). Moreover, little is known about the processes by which teacher characteristics might influence student achievement. Less is known about how teachers develop desirable characteristics during teacher preparation (Johnson & Birkeland, 2008; Wilson et al., 2002; Wilson et al., 2009).

The current lack of knowledge of effective ways to help teachers learn content knowledge for teaching impedes the ability of teacher education policies to make a difference in the classroom. Between 2002 and 2007, over 1 billion federal dollars were spent on Math-Science Partnerships, a professional development funding mechanism designed to increase the content knowledge of mathematics and science teachers; yet these programs and other professional development efforts over the last decade in the United States have unfortunately had at best modest effects on increasing teachers' mathematical knowledge for teaching (Hill, 2011).

One promising avenue forward is to explore and potentially leverage the interdependence of teachers' content knowledge and beliefs about teaching. If teachers' content knowledge and beliefs are truly interdependent, then focusing on content knowledge in isolation may not be enough. If mathematics education in the United States is to improve by improving the quality and effectiveness of teaching, then a broader, more comprehensive approach to teachers' professional growth may be warranted. This dissertation improves on previous research that has focused on teachers' content knowledge by taking seriously the possibility that change in

knowledge and beliefs are related. Each study in the dissertation examined aspects of teachers' productive disposition for teaching together with teachers' content knowledge and aims to uncover relationships between the two.

Teachers' experience with children's mathematical thinking in the course of professional experience and during student teaching is an important theme in the research on teachers' knowledge and beliefs about mathematics and learning, and a key marker of quality for field experiences. A randomized experiment (Philipp et al., 2007) compared the change over a semester-long mathematics content course in the beliefs and content knowledge of prospective teachers who were assigned to guided experiences that focused on children's mathematical thinking with that of prospective teachers assigned to clinical experiences that lacked such a focus. The authors found significant differences between the groups and called for prospective research to investigate mathematical knowledge for teaching (MKT) as an outcome. Silverman and Thompson (2008) have also argued that teachers' experience and knowledge of children's thinking is critical for developing MKT.

Each study in this dissertation was framed in relation to research and theory suggesting the importance of teachers' experience with children's mathematical thinking. To the extent that professional experience can be viewed as a "treatment" that promotes the growth of mathematical proficiency for teaching, children's mathematical thinking is the active ingredient. This dissertation addresses the field's critical lack of knowledge about how teachers develop mathematical proficiency for teaching by exploring the concurrent development of MKT and teaching self-efficacy (TSE) on the job across a variety of school contexts and by extending experimental work (Phillip et al., 2007) on how MKT and beliefs about mathematics and teaching mathematics develop during student teaching.

Dissertation Overview

The three studies that comprise this dissertation share an overarching theoretical framework built from research on and theorizing about teacher knowledge and beliefs in the field of mathematics education and from Albert Bandura's (1986, 1997) social cognitive learning theory. Accounts of mathematics teacher knowledge and beliefs transcend the discipline of mathematics and emphasize the role of children's mathematical thinking. Social cognitive theory highlights the role of beliefs about oneself in human behavior and learning. It also highlights the reciprocal interaction of behavior, individual cognitive and affective factors, and the social environment in human functioning. In particular, the human capabilities of self-reflection and self-regulation allow people to make sense of past experiences and adjust their actions in the future. All three studies reported in this dissertation investigated theoretically informed hypotheses about how teachers develop mathematical proficiency for teaching.

I framed all three studies in the context of recent policy for teacher education. The first study (Texas) explored a feature of some alternative routes to teaching certification—early entry. Early entry teachers begin teaching full-time before completing the requirements for full licensure; for these teachers, the learning associated with student teaching happens after they begin working if at all. The second study (Georgia) focused on induction, mentoring, and professional development of beginning teachers. The third study (United States) examined student teaching, the period of professional experience in schools that is typically the culmination of traditional teacher preparation routes.

Chapters 3, 4, and 5 report on the three dissertation studies. In Chapter 6, the final chapter, I discuss results from the findings of each study and from across the set of findings taken together. I also discuss the implications of those results for policy concerning certification

routes, the professional development of beginning teachers, and student teaching. I then discuss the theoretical implications of the results. In particular, I take up questions related to the nature of teacher's content knowledge and beliefs and their measures in light of the findings. Finally, I discuss the implications of the findings for future research, especially for the design of interventions positioned within student teaching or targeting the professional development of beginning teachers.

CHAPTER 2

FRAMEWORK AND LITERATURE REVIEW

In this chapter, I define how the term *multiplicative reasoning* is used in the present dissertation. I next discuss mathematical proficiency for teaching and summarize research on teachers' knowledge and beliefs about teaching and about mathematics and past work developing measures of teachers' knowledge and beliefs. In the last section of the chapter, I discuss teaching experience and summarize what is known and what remains to be discovered about how teachers' mathematical proficiency changes during student teaching, induction, and early years on the job.

Multiplicative Reasoning

Two of the studies in this dissertation used measures of teacher knowledge that focus on the content domain of multiplicative reasoning. In this section, I define *multiplicative reasoning* and summarize key findings from the mathematics education research on the multiplicative reasoning of Grades K–12 students and their teachers. The discussion of multiplicative reasoning in this section lays the foundation for the next section, in which I examine existing measures of content knowledge for teaching, including measures that tap teachers' content knowledge for teaching multiplicative reasoning topics.

Defining Multiplicative Reasoning

In this subsection, I will discuss both what is *multiplicative* and what entails *reasoning* in the content domain of multiplicative reasoning. I begin with reasoning. To speak of a knowledge

domain as a kind of reasoning means that the knowledge signified goes beyond propositional knowledge; the domain of multiplicative reasoning cannot be written down as a list of facts.

For me, *reasoning* is of two kinds, both of which build on the notion that reasoning is fundamentally about argumentation made up of claims that are logically supported by warrants. The first kind is *formal reasoning* by which I mean the discursive activity involving an abstract linguistic system that is arbitrated by mathematicians. This kind of reasoning involves standard definitions and theorems, deductive and inductive practices of argumentation, proofs, and methods of validating proofs. The second kind of reasoning that has particular relevance for defining multiplicative reasoning is *quantitative reasoning* (Thompson, 1994); it is also mathematical, but it is concrete in the sense that its referents are perceptible to the senses or composed of referents that are perceptible to the senses. According to Smith and Thompson (2008), *quantities* arise when people conceptualize aspects or attributes of objects, events, or situations as being *measurable* (such as distance or speed). *Quantification* is the mental act of assigning a value to the aspect or attribute so conceived. The *measure* of a quantity defines a unit and a process for assigning a number that represents the (proportional) relationship between a particular value of the quantity and the unit (Thompson, 2011). Therefore, quantitative reasoning is by definition proportional and multiplicative.

I turn next to the *multiplicative* aspect of multiplicative reasoning. The content domain of multiplicative reasoning cannot be easily defined in reference to school curriculum—multiplicative topics appear as early as second or third grade (e.g., whole number multiplication and division by grouping and sharing) and continue to be developed in secondary and even post-secondary coursework (e.g., similarity in Euclidean geometry and the derivative in calculus). Instead, multiplicative reasoning is defined by appealing to abstract mathematical

relationships—multiplicative structures—that are common to a wide range of curricular topics and applications (including those in science, technology, and engineering) and circumscribe a large set of problems (Vergnaud, 1983, 1988).

Vergnaud (1983, 1988) discussed three types of multiplicative structures. In this dissertation, I am not concerned with the whole domain of multiplicative reasoning, but only with an elementary subset. I therefore restrict my definition of (elementary) multiplicative reasoning to the first mathematical structure discussed by Vergnaud (1983), the *isomorphism of measures*, "a simple direct proportion between two measure-spaces" (p. 129). Vergnaud excluded inverse proportions from this structure, but he included multiplication, sharing and measurement division, and "rule-of-three" problems (p. 132). Thus, this multiplicative structure captures multiplication and division problems (including fraction arithmetic) and the fraction and ratio comparison problems and missing value problems that have traditionally defined the domain of ratio and proportion (Lamon, 2007; Tourniaire & Pulos, 1985).

Elementary multiplicative reasoning means supplying reasons in support of claims made about problem situations classified as having the isomorphism of measures multiplicative structure, and consists of the ability to form multiplicative comparisons between two quantities. (A multiplicative comparison is a comparison of two quantities *A* and *B* that answers the question, How many times as much (many) is *A* than *B*?). Because I do not take on advanced multiplicative reasoning topics, I will use the shorthand *multiplicative reasoning* hereon to refer to elementary multiplicative reasoning as it has just been defined. This definition of (elementary) multiplicative reasoning very closely follows the one proposed by Lamon (2007) for proportional reasoning.

> *Proportional reasoning* means supplying reasons in support of claims made about the structural relationships among four quantities (say *a, b, c, d*) in a context simultaneously

> involving covariance of quantities and invariance of ratios *or products* [italics added]; this would consist of the ability to discern a multiplicative relationship between two quantities as well as the ability to extend the same relations to other pairs of quantities. (p. 638)

I have two objections to Lamon's definition of proportional reasoning that led me to define multiplicative reasoning as I have.

My first objection is that Lamon's definition seems too broad. I agree that proportional reasoning is "a long-term process that is not easily segmented" (Lamon, 2007, p. 637). For consistency, if problems with invariant products are included in the set of situations that may elicit what's to be called "proportional reasoning," then many other problems which are mathematically similar should be included as well, including the full range of problems that constitute the *product of measures* multiplicative structure and the *multiple proportion* multiplicative structure discussed by Vergnaud (1988). Yet Lamon made an effort to distinguish between *proportional reasoning* and the broader notion *understanding of proportionality* that "develops only as one studies higher mathematics and science" (p. 640). For Grades K–12 school mathematics, the most important multiplicative structure is the isomorphism of measures (Behr et al., 1992). In my view, problems with invariant products (and even the so-called simple inverse proportion problems) are the provenance of this more advanced understanding of proportionality. Invariant product problems and invariant ratio or quotient problems are more different than alike; I will return to this point later.

My second objection to Lamon's (2007) definition is that the term "proportional reasoning" is too narrow for all that is captured by Lamon's definition. Vergaud's (1988) analysis of the *isomorphism of measures* multiplicative structure suggests that multiplication and division are special cases of the reasoning Lamon described. In particular, if any one of the quantities (*a, b, c, d*) in her definition has the value 1, then any problems that can be posed can

be interpreted as multiplication or division problem. The reasoning that Lamon's definition of proportional reasoning includes is more broadly multiplicative than the term proportional reasoning may suggest.

The revelation that proportional reasoning—challenging for many secondary students—has the same mathematical structure as whole number multiplication and division suggests that it may be possible to bootstrap students' (and teachers') knowledge of whole number multiplication to support their understanding of other aspects of elementary multiplicative reasoning (A. Izsák, personal communication, July 20, 2011). In Appendix A, I present a conceptual analysis of multiplicative reasoning that informs the instructional trajectory currently used in the Numbers and Operations content courses for middle grades and secondary teachers at the University of Georgia (Jacobson & Izsák, 2012a). In that analysis, I begin with whole number multiplication defined as grouping, then define fractions in terms of multiplication, use fractions and multiplication to describe two types of division, and finally demonstrate how proportions are a generalization of division. I also show how invariant product (inverse proportion) problems have mathematically distinct multiplicative structures, a conclusion that supports recent empirical work suggesting that direct and inverse proportion problems are psychologically distinct (e.g., Jacobson & Izsák, 2012b, Izsák & Jacobson, 2013).

Children's and Teachers' Multiplicative Reasoning

For decades, extensive research has been conducted on students' multiplicative reasoning (see Behr et al., 1992, and Lamon, 2007, for reviews). Although problems solved through multiplicative reasoning have a common structure, the domain itself is extremely complex. Research on multiplicative reasoning has focused on children's reasoning about multiplication and division of whole numbers, fractions, and rational numbers; on children's reasoning with

ratios including their understanding of ratio and fraction comparison and of ratio as a measure of an intensive quantity (see Appendix A); and on children's reasoning with proportions. Compared with the body of research on children's multiplicative reasoning, little is known about teachers' multiplicative reasoning. Many studies have found that some teachers have difficulty with the content their students should be learning, and that finding suggests that there may be parallels between children's and teachers' knowledge of the domain (e.g., Cramer, Post, & Currier, 1993; Harel & Behr, 1995; Post, Harel, Behr, & Lesh, 1991; Sowder, Philipp, Armstrong, & Schappelle, 1998).

Multiplication, division, and fractions. Researchers have used conceptual analysis to identify subconstructs of rational number including quotient, measure, ratio number, and multiplicative operator (Kieren, 1988, 1993; Behr et al., 1992). A fifth subconstruct called part-whole relationships was identified by Behr, Wachsmuth, Post, and Lesh (1984). These subconstructs are related to various ways of reasoning about multiplication and division of whole numbers and fractions and are also relevant for reasoning about proportional quantities. Lamon (2007) argued that empirical research focused on children's thinking may hold more promise for mathematics education than theoretical analyses. Some of this work has identified the importance of partitioning (equal-sharing) in children's understanding of multiplication and division with whole numbers and fractions (e.g., Confrey, 1994; Confrey & Smith, 1994, 1995; Empson, 1999; Empson, Junk, Dominguez, & Turner, 2005; Empson & Turner, 2006). Another line of work that has also dealt with students' partitioning has additionally shown the important role that children's understanding of units plays in how they make sense of multiplicative relationships (Hackenberg, 2007, Hackenberg, 2010, Hackenberg & Tillema, 2009; Steffe, 1988, 1993, 2001, 2003, 2004).

A number of studies have documented teachers' limited understanding of multiplication and division of fractions and decimals. One overarching finding is that teachers' have difficulty forming accurate verbal or graphical representations of the operations of multiplication and division and their results in the context of problem situations—a hallmark of quantitative reasoning (e.g., Ball, Lubienski, & Mewborn, 2001; Eisenhart et al., 1993; Izsák, 2008; Izsák, Tillema, & Tunç-Pekkan, 2008; Sowder et al., 1998). More evidence that teachers lack proficiency with quantitative reasoning comes from studies that described teachers incorrectly solving problems with situations that call for multiplying decimals (e.g., Graeber & Tirosh, 1988; Graeber, Tirosh, & Glover, 1989; Harel & Behr, 1995). Other studies reported the challenges U.S. teachers experience when explaining the operation of division in problem situations (e.g., Armstrong & Bezuk, 1995; Ball, 1990; Borko et al., 1992; Graeber & Tirosh, 1988; Izsák, Jacobson, de Araujo, & Orrill, 2012; Ma, 1999; Rizvi & Lawson, 2007; Simon 1993). Another over-arching finding from these studies is that teachers frequently rely on additive conceptions of multiplication and division (i.e., repeated addition and subtraction). Teachers may lack multiplicative conceptions of division such as those contained in the following quantitative division questions: How many are in one group? (sharing division); How many times as many/much? (measurement division).

Ratios and fractions. Forming a ratio is a kind of quantitative reasoning that requires the coordination of two quantities using a multiplicative comparison. Many researchers have found this coordination is difficult for children. One limitation arises from students focusing on change in a single quantity and ignoring others which precludes the formation of a ratio (Harel, Behr, Lest, & Post, 1994). Children can also form inappropriate additive comparisons instead of multiplicative comparisons (e.g., Hart, 1981, 1988; Karplus, Pulos, & Stage, 1983). The ratio

between two quantities forms a new quantity, and the ratio in fraction form can be understood as a rational number that conveys the value of the new quantity. This concept can be challenging for children to understand; researchers have documented the difficulty that students have reasoning about ratios as measures of speed (e.g., Thompson, 1994; Thompson & Thompson, 1994), ratios as measures of the slope of a ramp (e.g., Lobato & Siebert, 2002; Lobato & Thanheiser, 2002), and ratios as a measure of the intensity of a taste (e.g., Harel et al., 1994).

Several studies show that teachers can have difficulty understanding ratios and fractions. In a large-scale study of kindergarten teachers' knowledge of proportional reasoning that used Lamon's (2007) framework, Pitta-Pantazi and Christou (2011) found that although the teachers did relatively well on traditional missing value and comparison problems, they did poorly on items that tapped other conceptions of fraction and ratio, including those intuitive conceptions that research has shown many children bring with them into the classroom. In other studies, prospective elementary teachers struggled to make sense of ratios that expressed the measure of quantity like speed (the ratio of distance and time), the color or taste of a mixture, and steepness (e.g., Akar, 2010; Simon & Blume, 1994; Thompson & Thompson, 1994).

Ratios and proportions. Conceptions of ratios and the ability to form multiplicative comparisons between quantities are foundational elements of children's multiplicative reasoning about situations where covarying quantities maintain an invariant (constant) quotient, yet instruction can lead to reliance on poorly understood algorithms (Lamon, 2007).). Many researchers have found that students apply methods for solving problems with proportional relationships to problems that have a constant difference (i.e., $a - b = c - d$), constant product or inverse proportion (i.e., $ab = cd$), and other nonproportional relationships (e.g., Lamon, 2007, Van Dooren, De Bock, Janssens, & Verschaffel, 2008). Prior to instruction, students use a wide

range of appropriate strategies for reasoning with ratios, including forming a *composed unit* by joining two quantities that can then be iterated and partitioned to solve problems involving proportional relationships (Confrey, 1994; Lamon, 1995). Students can also solve problems involving proportional relationships using multiplicative comparison (Kaput & Maxwell-West, 1994). Lobato and Ellis (2010) argued that more sophisticated proportional reasoning such as that involving multiplicative comparison can arise from composed unit reasoning as children reflect upon the number of groups they create through successive iteration or partitioning operations. Thompson (1992) has argued that the most sophisticated proportional reasoning involves *rate*, the "reflected abstraction of constant ratio" (p. 7). With this conception, children recognize that a ratio represents an equivalence class of ratios; it is clear that rational numbers as defined in abstract mathematical terms are intimately connected to this form of proportional reasoning.

Teachers share many of the misconception that students have about ratio and proportions (e.g., Cramer et al., 1993; Harel & Behr, 1995; Simon & Blume, 1994; Sowder et al., 1998), and several studies show that teachers can rely on formal algorithms like cross multiplication even if their use is not appropriate (e.g., Fisher, 1988; Harel & Behr, 1995; Orrill & Brown, 2012; Riley, 2010). Just as students do, teachers make additive comparisons when multiplicative comparisons are appropriate (e.g., Canada, Gilbert, & Adolphson, 2010; Lim, 2009; Son, 2010). Teachers can have difficulty coordinating the covarying quantities in a proportion (e.g., Orrill & Brown, 2012). On the other hand, teachers make use of knowledge resources such as strategic multiplication and division when solving proportional reasoning problems that students do not use (Lobato, Orill, & Jacobson, under review).

Teachers learning and multiplicative reasoning. Increasing teachers' content knowledge in the domain of multiplicative reasoning is challenging work. There are examples of successful professional development that resulted in measurable gains in teachers' ability to reason about division, fractions, and proportional relationships (e.g, Ben-Chaim, Keret, & Ilany, 2007; Orrill & Brown, 2012; Sowder et al., 1998). Other cases are discouraging in that carefully designed interventions seemed to have little effect on teachers' knowledge (e.g., Garet et al., 2011, Jacobson & Izsák, 2012b). In the next section, I describe how teachers' knowledge of the domain of multiplicative reasoning is related to the mathematical knowledge they might use for teaching topics in this domain.

Mathematical Proficiency for Teaching

Kilpatrick et al. (2001) argued in the National Research Council report *Adding It Up* that teachers' mathematical proficiency for teaching depended on their knowledge of and beliefs about mathematics and learning. A compelling set of prior and subsequent results substantiates that claim. Teachers need to know the content they teach, and other kinds of teacher knowledge that are consequential for student learning have long been recognized. Historically, as measures of mathematics teachers' knowledge have increasingly focused on mathematical knowledge that is used in practice (rather than advanced disciplinary knowledge), the strength of the observed relationship with student achievement has increased (Hill, Sleep, Lewis, & Ball, 2007). Several recent measures of teacher knowledge have been framed in light of Shulman's (1986) notion of pedagogical content knowledge, a construct I discuss at length in the first part of this section. In the second part, I discuss teachers' beliefs about mathematics learning and teaching and how those beliefs affect teachers' instruction, including their use of knowledge in the classroom. In

the last part, I review literature describing the relationship between knowledge and beliefs and the rationale for studying them simultaneously.

Content Knowledge for Teaching and Its Measures

Little is known about the content knowledge that teachers need for teaching in part because most of the studies of teachers' knowledge have relied on crude proxies such as the number of content courses taken or degrees (e.g., Ball, Lubienski, & Mewborn, 2001; National Mathematics Advisory Panel, 2008). Early attempts to link teachers' mathematical knowledge with student performance led Begle (1979) to conclude that teacher's knowledge and effectiveness could not be linked. Yet in the last decade, large-scale studies using more sharply focused instruments have found evidence of the expected relationships between teacher knowledge and student achievement (Baumert et al., 2010; Hill, Rowan, & Ball, 2005; Tchoshanov, 2010).

These new instruments share a focus on the content knowledge that teachers' arguably use in practice, but the approach used by each research group was different. Ball, Thames, and Phelps (2008) proposed a framework for content knowledge for teaching ("subject-matter–specific professional knowledge", p. 389) that can be used to classify each approach. Tchoshanov (2010) designed the Teacher Content Knowledge Survey (TCKS) instrument for teachers with content objectives that were "closely aligned with corresponding objectives in state-mandated standardized tests for students" (p. 148). It included three subconstructs for different types of knowledge at different cognitive levels: procedural, conceptual, and abstract (e.g., explaining and proving). Only the conceptual knowledge sub-construct had a statistically significant relationship with student achievement, but this correlation was substantial ($r = .26$).

In the second study, Hill et al. (2005) used the Learning Mathematics for Teaching (LMT) instrument that also included items about the content knowledge that teachers might teach; for example, an item about evaluating exponents (p. 387). This kind of knowledge (like that assessed on Tchoshanov's instrument) is expected of educated adults, and Ball et al. (2008) called it *common content knowledge* (CCK). The measure used by Hill and her colleagues included items of another kind that were more the special province of teachers; for example, an item that asked whether or not a student strategy was valid (p. 388). The authors argued that these items assessed a kind of mathematical knowledge that teachers but few other adults or professionals would possess. Ball et al. (2008) called this kind of content knowledge for teaching *specialized content knowledge* (SCK). SCK goes beyond the content of instruction and informs teachers' explanations of why standard algorithms work and not just how to use them. It allows teachers to make sense of students' mathematical ideas, reasoning, and nonstandard strategies (Ball et al., 2008). Common content knowledge and specialized content knowledge items were modeled together on the same, one-dimensional scale using Item Response Theory (IRT). The major contribution of the study by Hill et al. was to provide evidence that teachers' content knowledge matters for student learning; the estimated effect size of teachers' content knowledge on student achievement was "in league with the effects of student background characteristics" (p. 396).

The third project called the Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy (COACTIV, Baumert et al., 2010) created a measure of teacher knowledge that had two distinct dimensions: content knowledge and *pedagogical content knowledge* (PCK). The COACTIV content knowledge instrument was similar to the second and third subconstruct of the instrument used by

Tchoshanov (2010) in that it focused on teachers' conceptual understanding of mathematical topics in the curriculum, and "all items required complex mathematical argumentation or proofs" (Baumert et al. p. 148). The COACTIVE PCK measure had three dimensions: identification of multiple solutions for *tasks*, recognition of *students'* thinking, and knowledge of representations and explanations for *instruction.* Baumert et al. (2010) found that PCK had greater power than content knowledge to predict student achievement, and was associated with higher student cognitive activation during instruction and greater individual learning support (p. 164). One contribution of the Baumert et al. study was in creating measures of two dimensions of teacher knowledge that were empirically distinct. Hill, Schilling, and Ball (2004) attempted to measure aspects of pedagogical content knowledge such as teachers' knowledge of content and students, but results from their factor analysis did not support the theoretical claim that this knowledge was a distinct subdomain from the mixture of CCK and SCK items that were used to form a unidimensional IRT scale.

All three of the instruments I have discussed are examples of measures of *mathematical knowledge for teaching* (MKT) in the Ball et al. (2008) framework. MKT includes pure content knowledge (CCK and SCK) as well as a mixture of content knowledge and pedagogical knowledge called pedagogical content knowledge (PCK). The term PCK was introduced by Lee Shulman (1986) and captured a provocative idea that has energized research on teachers' professional knowledge: Some of the knowledge that teachers use when teaching is a transformed combination of two distinct knowledge domains—knowledge of the content they are teaching and knowledge of general pedagogy. Shulman called for assessments that "could distinguish between a biology major and a biology teacher, and in a pedagogically relevant and important way" (p. 10). The Baumert et al. (2010) group succeeded in meeting this challenge,

and measures of MKT stemming from Hill and Ball's pioneering work are supported by a validity argument that includes strong relationships with student achievement and instructional quality (Hill et al., 2008; Hill et al., 2005; Hill, Kapitula, & Umland, 2011).

Part of the confusion around the use of the terms PCK and MKT may stem from the fact that the Hill et al. (2005) paper is one of the most widely cited papers on the significance of the MKT construct but it reported results from an MKT instrument that did not include any items on teachers' knowledge of students (p. 387) and thus did not include any PCK items as classified by Ball et al. (2008). Baumert et al. (2010) classified SCK as a component of PCK, and argued that that the Hill et al. instrument tapped PCK because it included "mathematical knowledge related to the instructional process" (p. 141). Other mathematics education researchers have also conceptualized the domain of content knowledge for teaching in different ways. For example, many descriptions of MKT have focused on the "deep" or "profound" understanding of mathematics that mathematics teachers need to support instruction that promotes student understanding (e.g., Ball, 1993; Ma, 1999; Simon, 2006).

More recently, Schmidt et al. (2007) and Tatto et al. (2008) conceptualized teachers' mathematics content knowledge (MCK) as advanced high school knowledge (rather than as deep conceptual understanding of school mathematics) for the Teacher Education and Development Survey in Mathematics (TEDS-M), an international comparison study conducted by the International Association for the Evaluation of Educational Achievement. The TEDS-M study also included a measure of mathematics PCK (MPCK), and the framework for that measure had three components: mathematics curricular knowledge, knowledge of planning for mathematics teaching and learning, and enacting mathematics for teaching and learning (Tatto et al., 2012). Other research groups have created measures of MKT or PCK for a variety of purposes,

including assessing teacher education programs (e.g., Diagnostic Teacher Assessment in Mathematics and Science [DTAMS]; Saderholm, Ronau, Brown & Collins, 2010) or assessing specific professional development interventions (e.g., SimCalc; Shechtman, Roschelle, Haertel, & Knudsen, 2010). The DTAMS project assessed four types of knowledge. The first three mapped closely to Tchoshanov's (2010) types: content knowledge that is procedural (Type 1), conceptual (Type 2), and involving higher-order thinking (Type 3). The fourth type of knowledge assessed by DTAMS was PCK; it included "identifying and correcting student misconceptions and errors, creating analogies and examples to explain procedures and phenomena, and helping students make connections across mathematical concepts and ideas" (Saderholm et al., 2010, p. 182). In the SimCalc instrument (Shechtman et al., 2010), MKT was defined as "essentially mathematical knowledge, but a specialized type that teachers need to make sense of students' mathematical work" (p. 328).

The field has also seen recent innovation in test design. Kersting, Givvin, Thompson, Santagata, and Stigler (2012) designed the Classroom Video Analysis (CVA) measure of teacher knowledge that used a rubric to assess teachers written responses to video clips of instruction. The Diagnosing Teachers Multiplicative Reasoning (DTMR) project (Bradshaw, Izsák, Templin, & Jacobson, under review) has developed two paper-and-pencil instruments for measuring middle grades teachers' knowledge that use Diagnostic Classification Models (DCMs), an emerging family of psychometric models. The first assessment is intended to measure aspects of multiplicative reasoning critical for multiplication and division of fractions; the second assessment is intended to measure core aspects of proportional reasoning. In terms of the framework for mathematical knowledge for teaching, the DTMR tests emphasize SCK and conceptual understanding of the content. The knowledge assessed includes teachers' reasoning

with quantities such as lengths, areas, and volumes rather than using computational procedures, and it assessed teachers' understanding fraction arithmetic and proportional relationships in the context of problem situations and drawn models (e.g., number lines and rectangular areas).

Table 2

Content Knowledge for Teaching Instruments, Constructs, Subconstructs, and Categories Classified by Three Kinds of MKT

Project or Instrument	Kinds of MKT		
	Conceptual (or advanced) knowledge of the content taught	Knowledge to understand or appraise students' responses and reasoning (mathematical thinking)	Knowledge about the mathematical and instructional entailments of tasks and representations
COACTIV	• Content knowledge • PCK-Tasks	• PCK-Students	• PCK-Instruction
DTAMS	• Type 2 (conceptual) • Type 3 (higher-order)	• Type 4 (PCK)	• Type 4 (PCK)
DTMR	Implicitly assessed	Not assessed	Explicitly assessed
LMT	• CCK items • Implicitly assessed in SCK items	• SCK items • Knowledge of content & students (not assessed)	• SCK items
SimCalc	Explicitly assessed	Explicitly assessed	Explicitly assessed
TCKS	• Type 2 (conceptual) • Type 3 (abstract)	Not assessed	Not assessed
CVA	Not assessed	Explicitly assessed	Not assessed
TEDS-M	• MCK	• MPCK	• MPCK

Note. *Explicitly assessed* means that some aspect of the framework for the construct can be mapped to this category of MKT; *implicitly assessed* means that descriptions of the instrument invoked the argument that some level of content knowledge is required to answer PCK items so PCK items implicitly assess content knowledge.

The Ball et al. (2008) framework for MKT is a useful lens for understanding what has been accomplished over the last decade in measuring mathematics teachers' professional knowledge. The framework also provided promising avenues for future research that are outside the scope of this discussion (e.g., horizon knowledge, p. 403). As I have described, the terminology and definitions suggested by the framework have not proved canonical, and using those terms and definitions when speaking across projects invites misunderstanding.

Looking across projects that had successfully measured constructs, components, or categories of MKT, I noticed three overarching kinds of MKT: (a) conceptual (e.g., TCKS: Type 2) or advanced (e.g., TEDS-M) knowledge of the content taught, (b) knowledge to understand or appraise students' responses and reasoning (i.e., their mathematical thinking; e.g., COACTV: PCK-Students), and (c) knowledge about the instructional use of mathematical tasks (e.g., DTMR). The last two kinds of MKT are directly related to specific kinds of work (pedagogical tasks) that teachers must do in the classroom: understand and appraise student responses and select and use tasks and representations for instruction (Ball et al., 2008). Table 2 shows how I classified various instruments, constructs, subconstructs, and categories of MKT by these three kinds.

The classification in Table 2 makes it clear how the terms one might expect to have common meanings (e.g., PCK, SCK, CCK) do not coincide across projects. For example, items that formed the COACTIV: PCK-Tasks subconstruct required teachers to provide multiple representations and solutions. Teachers with this capacity clearly would have more opportunities for providing rich mathematical explanations and better instructional quality, but those activities require something beyond the ability to generate multiple solutions and explanations. On its own, this category of COACTIV seems quite similar to the characterization of SCK items used

on the LMT instrument: items that assess mathematical capacities teachers need that other adults and professionals may not. In other projects, no effort was made to classify the instrument or parts of the instrument with subconstructs or categories of MKT and PCK or to distinguish MKT and PCK (e.g., CVA, TEDS-M, and SimCalc). Another striking observation that can be made from this classification is the range in choices that projects have made with respect to framing "pure" content knowledge: It is assessed explicitly (and, for TCKS, is the entirety of the domain); it is assessed implicitly by other kinds of MKT items because PCK logically depends on some minimum amount of content knowledge (e.g., DTMR); or it is not assessed (e.g., CVA). Given the progress that these groups have made by creating measures that have provided some of the first evidence of how content knowledge for teaching (conceptualized and measured in the various ways that have been described) is related to student learning, researchers efforts would be advised to move future research in this area towards more synthesis.

Productive Disposition for Teaching and Measures of Teachers' Beliefs

Knowledge is only one component of mathematical proficiency for teaching. Teachers' beliefs and affect, orientations and dispositions have important consequences for the work of teaching as well. Work in psychology and in mathematics education over the last several decades also provides a foundation for operationalizing the notion of productive disposition for teaching mathematics. I repeat the definition from the first chapter for clarity: *Productive disposition for teaching mathematics* refers to mathematics teachers' orientation toward—and their related beliefs and attitudes about—the subject of mathematics, teaching and learning it, and their own professional growth. Two constructs that I used in the present studies concerned teachers' beliefs. The first construct, teachers' self-efficacy beliefs, has to do with what teachers believe about themselves as teachers. This belief is important because it is related to teachers'

motivation to teach. The second construct concerns teachers' beliefs about the nature of teaching and the discipline of mathematics. These beliefs are important because what teachers believe shapes how they select and pursue goals in the classroom (see Philipp, 2007, for a review).

Teaching self-efficacy (TSE) beliefs are a teacher's own judgments about her or his capability to teach and confidence that her or his instruction will affect student learning (Bandura, 1977; Pajares, 1992). The construct of TSE has been used extensively for several decades (e.g., Gibson & Dembo, 1984; Hoy & Woolfolk, 1993), and several measures of TSE exist (Tschannen-Moran & Hoy, 2001; Tschannen-Moran, Hoy, & Hoy, 1998). Under Bandura's (1986) social-cognitive theory, TSE beliefs determine teachers' "persistence when things do not go smoothly and their resilience in the face of setbacks" (Tschannen-Moran & Hoy, 2001, p. 784), and thus is clearly related to the productive disposition identified by the NRC (Kilpatrick et al., 2001).

In a comprehensive review, Tschannen-Moran et al. (1998) summarized the established associations between TSE and students' achievement and motivation. These authors concluded that TSE is likely to form early in teachers' careers and remain relatively difficult to change later on. The development of TSE is thus a crucial goal for programs preparing and supporting novice teachers. An extension of the Hill et al. (2007) argument for teacher knowledge measures that are close to the practice of teaching suggests that content-specific self-efficacy measures would be more highly predictive of student achievement in that content area than more general measures. Moreover, self-efficacy to teach may vary with the content taught. Bandura (1986) wrote that self-efficacy as such was too broad to be useful for research without narrowing one's attention to self-efficacy beliefs that are relevant to the specific situation or activity being researched.

The construct of self-efficacy plays a central role in the social-cognitive theory of psychology (Bandura, 1977, 1986, 1997). Self-efficacy influences "how much effort will be expended and how long it will be sustained in the face of obstacles and aversive experiences" (Bandura, 1977, p. 191), and is the key factor of human agency. Social-cognitive theory identifies four sources of self-efficacy:

> Enactive mastery experiences that serve as indicators of capability; vicarious experiences that alter efficacy beliefs through transmission of competencies and comparisons with the attainments of others; verbal persuasion and allied types of social influences that one possesses certain capabilities; and physiological and affective states from which people partly judge their capableness, strength, and vulnerability to dysfunction. (p. 79)

Reviewing measures of the sources of self-efficacy, Usher and Pajares (2008) critiqued the validity of measures that were not aligned with theory. Morris (2010), one of Usher's students, has done extensive work in selecting and validating measures of the sources of TSE, and has found compelling evidence that TSE and its sources are related as theory would predict.

Teachers' beliefs about mathematics and learning are also related to student achievement (see Philipp, 2007, for a review). For example, Fennema et al. (1996) found in a longitudinal study that teachers who have opportunities to understand children's mathematical thinking come to believe that learning mathematics is a process of inquiry involving active student participation and that students of teachers with these beliefs experience larger gains in achievement than students whose teachers believe that mathematics is a set of rules best learned by rote. Staub and Stern (2002) conducted a quasi-experimental study and found that the students of mathematics teachers with inquiry and active learning beliefs demonstrated greater achievement gains than students of teachers without those beliefs. Researchers on the TEDS-M project (Tatto et al., 2012) cited these promising results as a rationale for their use of similar measures to assess

teachers' beliefs about whether mathematics involves inquiry and whether mathematics students should be engaged in active learning.

Connections Between Knowledge and Belief

Outside of teaching situations, researchers have consistently found close relationships between knowledge and self-efficacy; for example, in studies of Grades K–12 mathematics students (see Pajares, 1996, for review). Although MKT and TSE are both correlated with student achievement, little work has been done exploring how these two consequential characteristics of teachers interact. Kilpatrick et al. (2001) argued that the components of knowledge and disposition in mathematical proficiency for teaching are interdependent; one cannot develop without the other. Thompson (1992), in her handbook chapter on mathematics teachers' beliefs and conceptions, argued that, "to look at research on mathematics teachers' beliefs and conceptions in isolation from research on mathematics teachers' knowledge will necessarily result in an incomplete picture" (p. 131); the inverse is certainly true as well.

It is plausible that MKT and TSE are interdependent, but little empirical evidence exists to support that hypothesis. In the only survey study that relates MKT with a disposition-related construct, Hill (2010) found that MKT is predicted by mathematical self-concept, one component of motivation that is likely correlated with TSE. The only qualitative study to investigate the interaction between teachers' MKT and teaching beliefs (Hill et al., 2008) discussed beliefs about mathematics and how students learn, but the researchers did not consider teachers' motivation or TSE. A promising, unexplored site for identifying the conditions and constraints governing teachers' simultaneous development of MKT and TSE is the changing practice of novice teachers.

Teachers' Professional Experience in Schools

Student Teaching

Recent policy recommendations for teacher preparation in the United States (e.g., National Research Council [NRC], 2010; National Council for the Accreditation of Schools [NCATE], 2010) have focused on the promise of school-based field experiences for producing desired outcomes for prospective teachers. Those recommendations echo an international consensus on the importance of clinical experiences for teacher education (e.g., Musset, 2010; Wang, Coleman, Coley, & Phelps, 2003) and an international trend over the last few decades among teacher educators of increased emphasis on clinical experiences (Maandag, Deinum, Hofman, & Buitink, 2007; Ronfeldt & Reininger, 2012). Policy recommendations call for increased field-based teacher education by scheduling earlier clinical experiences and extending their duration (Goodson, 1993; Villegas-Reimers, 2003).

At the same time, descriptive studies of clinical experiences in the United States suggest that they can be poorly aligned with teacher education program goals and that placements in schools can be haphazard, with little university oversight (Wilson et al., 2002). Some researchers argue that earlier or longer clinical experiences may be ineffective or even detrimental if the quality is poor; for example, by leading to beliefs about mathematics teaching and learning that are inconsistent with university course work (Zeichner & Gore, 1990).

Feedback from a supervisor or mentor during student teaching is critical, because novice teachers tend not to notice what might be of significance (such as how students are reasoning) during clinical experiences (e.g., Jacobs, Lamb, & Philipp, 2010). Research-based reports of high quality clinical experiences in the United States consistently describe feedback from supervising teachers as a critical component (Boyle-Baise, & McIntyre, 2008; Clift & Brady,

2005; Darling-Hammond & Bransford, 2007). Feedback has important consequences for the educative value of clinical experiences (Conderman, Morin, & Stephens, 2005; Fernandez & Erbilgin, 2009), and there is concurring evidence that student teacher mentoring has positive effects on prospective teachers' instructional practice (e.g., Murray, Nuttall, & Mitchell, 2008).

There is less empirical support for the recommendations for longer student teaching; the relevant research has been primarily descriptive and has frequently lacked adequate controls for selection bias. Some studies have provided evidence for positive effects of extended field experiences on teacher outcomes (e.g., Andrew, 1990; Andrew & Schwab, 1995; Silvernail & Costello, 1983). By contrast, large-scale studies (also lacking adequate controls) have compared teachers completing one versus two semesters of student teaching and found no difference in teaching self-efficacy beliefs (Chambers & Hardy, 2005; Spooner, Flowers, Lambert, & Algozzine, 2008). Only two studies of which I am aware estimated pseudo-causal effects for student teaching. Boyd, Grossman, Lankford, Loeb, and Wyckoff (2009) used a robust set of controls and found that estimates of the effect of no student teaching on teachers' value added to student achievement was unstable across models; Ronfeldt and Reininger (2012) used similar controls and concluded that the length of student teaching had no effect on teachers' preparedness to teach, but that the quality of student teaching had significant positive effects.

Teacher Development

The research literature on learning to teach is expansive but has been criticized for being fragmented and of uneven quality (Wideen et al., 1998). Until recently, researchers focused primarily on the effects of particular features of university teacher education programs or highly delimited professional development interventions. Only in the last decade or so has the focus of research shifted to analyzing learning to teach as a complex, situated process (Feiman-Nemser,

2001). Discussing the current state of professional development and promising directions, Borko (2004) emphasized the importance of a situated perspective for research on teachers' development that carefully attends to multiple contexts and influences. She also highlighted the consensus that professional development programs should focus on teachers' content knowledge and on developing their facility with student thinking.

Looking at teachers' first year of practice is not sufficient. Recent qualitative studies have suggested that effects of teacher preparation may appear during teachers' second and subsequent years of practice that were not apparent during the teachers' first years of practice (Ensor, 2001; Grossman et al., 2000; Peressini, Borko, Romagnano, Knuth, & Willis, 2004). Large gaps in the literature remain. For example, none of the research reviewed by Wideen et al. (1998) addressed teacher content knowledge or self-efficacy beliefs for teaching, and more than 10 years later, Charalambous (2009) observed that little is known about how novices learn MKT.

The idea that teachers learn from experience is certainly plausible; in fact, the proliferation of alternative certification routes may be due in no small part to policy arguments that rely on that assumption. Evidence exists that novice teachers become more effective in their first few years and (for mathematics teachers) is consistent with the hypothesis that novices learn MKT. In each study, however, alternative explanations of the results are not ruled out.

New teachers are not as effective as experienced teachers (e.g., Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), and a wide body of research has documented the positive association between experience and student achievement. The research has further suggested that the benefit associated with experience tapers off after 4 or 5 years (e.g., Clotfelter, Ladd, & Vigdor, 2006; Darling-Hammond, Berry, & Thoreson, 2001). These results were all possibly subject to cohort bias (varying employment conditions lead to hiring cohorts that differ in

characteristics that affect student achievement) and attrition bias (the teachers who leave and those who remain differ in characteristics that affect student achievement). Although inconclusive on the whole, research comparing teachers who follow traditional versus alternative routes appears to agree that most differences in effectiveness have equalized after 2 or 3 years of experience (Feistritzer & Haar, 2008). Some alternative routes, such as Teach for America, recruit very well-educated college graduates but have 80% attrition by the fourth year (Heilig & Jez, 2010), so the result that differences between traditional and alternative route teachers disappear as a consequence of experience may also be affected by attrition bias.

A much smaller set of studies has suggested a similar relationship between teacher experience and MKT. Hill (2007, 2010) found a significant linear relationship between experience and MKT among random national samples of elementary and middle grades teachers, but she hypothesized that cohort bias might be responsible for the observed effects. She also called for more research describing the content knowledge of alternative route teachers Using a covariance adjustment model of middle grades teachers' algebra MKT, Hill (2011) found that experience teaching algebra between the two waves of knowledge measurement had a large and significant effect ($d = .34$, $p < .001$) on algebra MKT after controlling for pretest scores. The results might have been biased, however, by large score increases among a relatively few teachers who were teaching algebra for the first time and by the recency effect documented in experimental psychology whereby the most recently seen items on a list are the easiest to recall (H. Hill, personal communication, August 29, 2011). In this case, having recently taught the material may have biased experienced teachers' scores upwards relative to those of teachers with similar knowledge who had not seen the material in over a year and were remembering "on the fly" while taking the knowledge survey. Humphrey, Wechsler, and Hough (2008) compared

participants in seven alternative certification programs and found that initial differences in MKT associated with the selectivity of undergraduate institution had disappeared by the end of the first year and that all teachers had made gains in MKT.

Recent results about professional development have not been as promising as the apparent effects of experience. Hill (2011) found only modest effects of professional development on teachers' MKT in a national sample of middle grades teachers. Garet et al. (2011) found in a large, well-designed research study that a mandatory professional development intervention for rational number did not have a statistically significant effect on teacher knowledge.

Research on how novice teachers' beliefs develop is rare, but several studies supported the claim that teacher preparation affects teachers' beliefs. Newton (2009) adapted a mathematics motivation instrument to examine prospective elementary teachers' motivation for solving mathematical problems involving fractions and found that a course on the conceptual understanding of elementary mathematics significantly increased their motivation for solving problems with fractions. Woolfolk and Hoy (1990) found that TSE increased during student teaching, and Darling-Hammond, Chung, and Frelow (2002) reported that traditional route teachers had significantly greater TSE than alternative route teachers with less preparation.

There has been some work on theoretical explanations for how teacher expertise might develop, but little has focused on teachers' professional knowledge. Berliner (1994) hypothesized that teachers develop expertise in stages but later observed (Berliner, 2001) that stage theories do not explicate the developmental process between stages that are of greatest interest to teacher educators. Silverman and Thompson (2008) presented a developmental account of learning MKT that fundamentally involved "decentering" (p. 502), or a teacher's

ability to set aside his or her own perspective and imagine moving through a particular mathematical terrain from the perspective of a student. This step, they argue, allows teachers to transform their own personal mathematical understanding into a powerful pedagogical resource. In a similar vein, Darling-Hammond (2000) observed:

> Developing the ability to see beyond one's own perspective, to put oneself in the shoes of the learner and to understand the meaning of that experience in terms of learning, is perhaps the most important role of universities in the preparation of teachers. (p. 170)

It is also a role that must not be neglected by professional development programs and other forms of support for inservice teachers, especially those aimed at increasing teachers' MKT.

Teachers' experience with children's mathematical thinking is an important theme in the research on teachers' knowledge and beliefs about mathematics and learning, and a key marker of quality for field experiences. A randomized experiment (Philipp et al., 2007) compared the change over a semester-long mathematics content course in the beliefs and content knowledge of prospective teachers who were assigned to guided experiences that focused on children's mathematical thinking with that of prospective teachers assigned to clinical experiences that lacked such a focus. A critical design feature was the early timing of the clinical experience to be concurrent with a content course. Philipp et al. (2007) hypothesized that experience with children's mathematical thinking would promote the prospective teachers' development of beliefs and content knowledge; they found significant differences between the groups with respect to changes in beliefs but no significant differences in changes in content knowledge. They did not use an instrument to measure PCK and called for future research to investigate measures of that construct.

CHAPTER 3

THE TEXAS STUDY

This study explores the relationship between professional experience and mathematical proficiency for teaching. Teachers certified for high school usually take more mathematics classes in college and may enjoy mathematics more than teachers certified for earlier grade levels. High school teachers therefore may have more knowledge resources and more motivation for explaining multiplicative reasoning topics than middle grades or elementary teachers. On the other hand, teachers frequently teach only a small portion of the curriculum, and one might reasonably expect an experienced fifth-grade teacher to have more mathematical proficiency for teaching fraction division than a similarly experienced second-grade teacher or high school geometry teacher. Past research has not addressed the extent to which experience teaching mathematics is conducive for developing mathematical knowledge for teaching.

Existing coarse-grained instruments assess a wide range of knowledge and beliefs, much of which specific teachers may rarely use in practice. In the Texas study, I narrowed the focus of instruments for measuring mathematical proficiency for teaching by selecting or adapting existing items to form new instruments that targeted the specific content domain of multiplicative reasoning. This sharp focus afforded a more meaningful analysis of the relationship between mathematical proficiency for teaching and features of teachers' experience, such as the grade level at which they teach.

The first purpose of the Texas study, therefore, was to assess the reliability of the adapted instruments of mathematical proficiency for teaching. Mathematical proficiency for teaching

was operationalized in the Texas study on two dimensions: mathematical knowledge for teaching and teaching self-efficacy beliefs. A separate instrument was developed for each of these dimensions of mathematical proficiency for teaching. In addition, supplemental instruments were adapted or used verbatim on the survey to inform the validity argument for the instrument of mathematical knowledge for teaching multiplicative reasoning and the instrument of self-efficacy beliefs for teaching multiplicative reasoning.

The second purpose of the Texas Study was to study how mathematical proficiency for teaching multiplicative reasoning varied in a sample of *practicing teachers* with diverse preparation experiences and who were working in a wide range of grade levels and school contexts. A key feature of this study was the opportunity to compare *early entry* teachers who begin teaching before completing the requirements for certification with those teachers who were certified before beginning to teach. Early entry status is of greater interest than a comparison between, say, alternative and traditionally certified teachers because the former is a specific feature that can be influenced directly by policy, whereas the latter distinction holds little meaning because of the great variation within and between alternative and traditional certification programs (Johnson & Birkland, 2008). My research in service of the second purpose of the study was essentially descriptive, and it laid the empirical groundwork for the hypotheses that were further explored in the longitudinal Georgia study (see Chapter 4) and in the study of U.S. student teaching (see Chapter 5).

The research questions for this study were as follows:

1. How valid and reliable are the content knowledge and teaching self-efficacy measures adapted for the domain of multiplicative reasoning?

2. How does mathematical proficiency for teaching multiplicative reasoning vary (a) with academic preparation, student teaching, and teaching experience and (b) by early entry status and across certification grade level?

Data and Methods

Participants

The analytic sample for this study comprised volunteers from among the participants in two studies of Grades K–12 mathematics teachers in Texas conducted by Michigan State University (see http://usteds.msu.edu). The two Texas studies were designed to supplement the U.S. portion of the international Teacher Education and Development Survey in Mathematics (TEDS-M; Tatto et al., 2012), which had sampled only teachers prepared in public institutions. Because of the large and growing number of teachers following alternative routes to certification in the United States, those studies were aimed to provide data on teacher education more broadly. The prevalence of teachers who follow alternative routes to certification varies widely by state, and Texas was a strategic choice because of the large percentage of new teachers in Texas who follow an alternative route. I asked the volunteers to complete the Teachers and Teaching Multiplicative Reasoning (TTMR) survey, which contained the focused measures of teachers' knowledge and beliefs I had developed, as well as various other questions about teachers' background, preparation, and experience. Data from the TTMR survey and from the U.S. TEDS-M supplemental studies were used in the study reported in this chapter.

The first U.S. TEDS-M supplemental study (TX-TEDS1) was conducted between June and October of 2010 and aimed to reach all Grades K–12 mathematics teachers in Texas who had been given initial certification between 2006 and 2010 to teach either in the elementary grades (generalist, early childhood, EC, to Grade 6) or in the middle and upper grades

(mathematics, Grades 4 to 8 and Grades 8 to 12). There were approximately 17,750 individuals contacted for the first survey, but only 1,015 completed it (a response rate of 5.7%). These individuals were invited to take the TTMR survey, and of the 166 who volunteered to do so, 106 responded to at least one question on the TTMR survey (64%).

The second study (TX-TEDS2) differed from TX-TEDS1 in two ways. First, it was a representative probability sample of the population of Grades K–12 mathematics teachers certified to teach in Texas between 2006 and 2010. Second, the teachers in the sample were offered a small honorarium to encourage participation. Approximately 8400 teachers were contacted and even though the sample was smaller, the number of participants that completed the survey was higher ($N = 1937$). The response rate for TX-TEDS2 was consequently higher than for TX-TEDS1 (23% versus 6%), and these participants volunteered to take the TTMR survey at a higher rate as well (27% versus 16%). Of 516 volunteers, 386 responded to at least one question on the TTMR survey (75%).

Instruments

The TTMR survey included an instrument of mathematical knowledge for teaching (MKT), a teaching self-efficacy (TSE) beliefs instrument (TSE Beliefs), and an instrument measuring the sources of TSE (TSE Sources). These instruments were adapted from existing instruments to focus on the domain of multiplicative reasoning. The TTMR survey also included several questions that addressed the background, preparation, and school context of the participants.

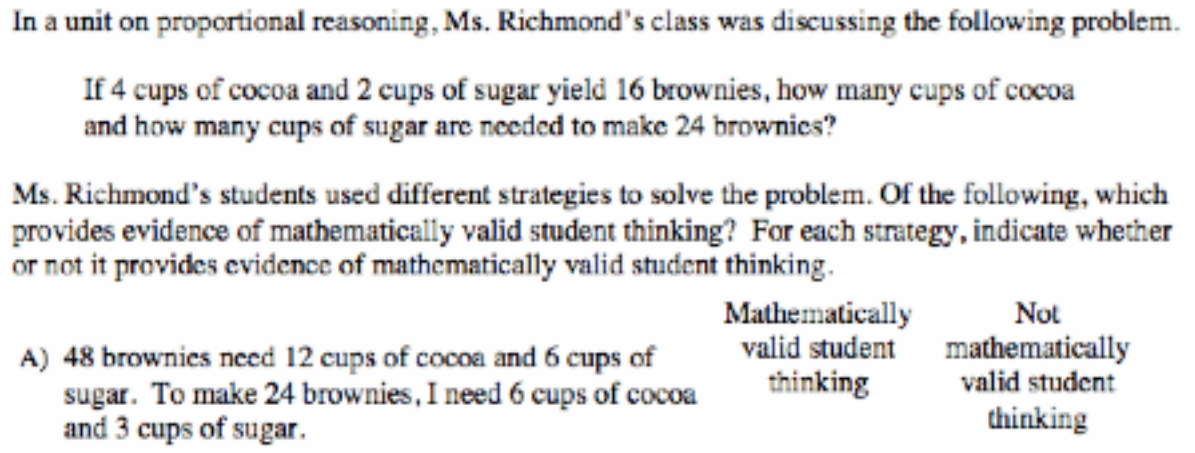
Mathematical Knowledge for Teaching instrument (MKT). I selected 23 items written for the Measures of Effective Teaching project (Bill & Melinda Gates Foundation, 2010) that focused on fractions, ratios, and proportions to form the MKT instrument used for this study.

The teacher knowledge instrument developed by the Measures of Effective Teaching project was similar in design to the LMT instrument (see Chapter 2). I also adapted 3 items from the Diagnosing Teachers Multiplicative Reasoning project (DTMR; Bradshaw, Izsák, Templin, Jacobson, under review).

The items that I selected explicitly addressed two kinds of MKT identified in Chapter 2: (a) understanding and evaluating students' mathematical thinking and (b) selecting and using tasks and representations. The third kind of MKT (conceptual knowledge of the content taught) was implicitly assessed by the selected items because such knowledge is prerequisite for answering them. The selected items could be further classified in three topical categories that make up the domain of multiplicative reasoning as defined for this study (fraction multiplication and division, fraction and ratio comparison, and proportional reasoning). These categories correspond to the categories I used to organize the review of students' and teachers' multiplicative reasoning in Chapter 2. Table 3 shows how the items were distributed in this cross-classification and it demonstrates the higher number (hence emphasis) of items on the MKT instrument that dealt with students' mathematical thinking. Figure 1 shows examples of the MKT items.

Teaching Self-Efficacy Beliefs (TSE Beliefs) instrument. The instrument for TSE Beliefs was adapted from measures for prospective science teachers (Enochs & Riggs, 1990; Roberts & Henson, 2000). The items measuring TSE Beliefs were modified to address the domain of multiplicative reasoning by replacing the word "science" with the phrase "topics involving fractions, ratios, and proportions." For example, the question "I usually do a poor job teaching science" became "I usually do a poor job teaching topics involving fractions, ratio, and proportion." Following Roberts and Henson (2000), the TSE Beliefs instrument had two factors:

personal teaching efficacy (PE) and knowledge efficacy (KE). Sample items for each of these factors are presented in Table 4. A central question in the instrument validation work described below was the evaluation of this two-factor structure for the TSE Beliefs instrument.



In a unit on proportional reasoning, Ms. Richmond's class was discussing the following problem.

If 4 cups of cocoa and 2 cups of sugar yield 16 brownies, how many cups of cocoa and how many cups of sugar are needed to make 24 brownies?

Ms. Richmond's students used different strategies to solve the problem. Of the following, which provides evidence of mathematically valid student thinking? For each strategy, indicate whether or not it provides evidence of mathematically valid student thinking.

	Mathematically valid student thinking	Not mathematically valid student thinking
A) 48 brownies need 12 cups of cocoa and 6 cups of sugar. To make 24 brownies, I need 6 cups of cocoa and 3 cups of sugar.		

(a)

In a unit introducing proportions, Mr. Hayes gave his students the following problem.

Sabrina's Flower Shop has 24 roses and 36 carnations. What is the ratio of the number of roses to the total number of roses and carnations at Sabrina's Flower Shop?

When the students finished, he wanted to give them a similar practice problem. Of the following, which asks students to do the most similar mathematical work and thinking as the problem above?

A) A spinner has 10 equal sections numbered 1 through 10. What are the odds that the spinner will land on a prime number when it is spun?

B) Nick's Team Sports Store has 24 basketballs and 60 footballs on display. What is the ratio of the number of basketballs to the number of footballs on display?

C) For breakfast during the month of April, Katherine ate cereal on 18 days and eggs on the other 12 days. What is the ratio of the number of days Katherine ate cereal for breakfast to the number of days in April?

D) All of these problems ask students to do the same mathematical work and thinking.

(b)

Figure 1. Sample (a) dichotomous response and (b) multiple-choice items used in the MKT instrument.

Table 3

Classification of MKT Items by Pedagogical Task, Content Topic, and Problem Type

	Pedagogical task		
Content topic	Understanding and appraising students' mathematical thinking	Selecting and using tasks and representations for instruction	Total
Proportional reasoning	5 – DR (e.g., Richmond in Fig. 1)	3 – DR	8
Fraction and ratio comparison	3 – MC 6 – DR	1 – MC (e.g., Hayes in Fig. 1)	10
Fraction multiplication and division	4 – DR	4 – DR	8
Total	18	8	26

Note. Cells give the number of items and the type in each category. Dichotomous response items (DR) had a common stem and asked participants to respond given two options (e.g., mathematically valid vs. not mathematically valid). Multiple-choice (MC) items had four options.

Sources of Teaching Self-Efficacy (TSE Sources) instrument. The TTMR survey included an instrument to assess the sources of teaching self-efficacy to support a validity argument based on Bandura's (1986, 1997) theoretical account of self-efficacy beliefs. Social cognitive theory stipulates four sources of self-efficacy beliefs: mastery experiences, vicarious experiences, social persuasions, and emotional and physiological states. *Mastery experiences* are individuals' interpretation and evaluation of their own competence and confidence after completing a task. Mastery experiences are the most powerful source of self-efficacy beliefs. Individuals also construct self-efficacy beliefs through *vicarious experiences* from peers or from their own memories of themselves. Vicarious experiences involve watching or remembering another's performance (including one's own past performance) and imagining oneself completing the task in a similar manner. *Social persuasions* from students, other teachers, and administrators provide another source for teachers to develop their self-efficacy beliefs. Finally,

individuals' *emotional and physiological states* before and after performing a task influence how they interpret their own competence. Anxiety, fatigue, and mood are all aspects of this fourth source of self-efficacy beliefs.

Table 4

Sample Items From the TSE Beliefs and the TSE Sources Instruments

	Factor	Example Item	α
TSE Beliefs	PE – personal efficacy	I am not sure I have the necessary skills to teach every topic involving fractions, ratios, or proportions.	.83 .77[a]
	KE – knowledge efficacy	I understand concepts involving fractions, ratios, and proportions well enough to be effective in teaching my students.	.88
TSE Sources	ME – mastery experience	I have succeeded at teaching topics involving fractions, ratios, and proportions even with the most challenging students.	.82
	VE – vicarious experience	When I am preparing to teach topics involving fractions, ratios, and proportions, I often try to visualize myself working through the most difficult teaching situations.	.58 .79[b]
	SP – social persuasion	My students have told me that I have taught them a great deal about topics involving fractions, ratios, and proportions.	.82
	EP – emotional & physiological states	I would be worried if I was asked to demonstrate how to teach a lesson that involved fractions, ratios or proportions.	.91

[a] One item was not used in the final PE scale because of differential additive effects across the certification grade level grouping; the second value for α indicates Cronbach's reliability coefficient after removing this item. See the Measurement Invariance section for more details.
[b] One item was not used in the final VE scale because of misfit, and the Spearman-Brown reliability coefficient was $\rho = .79$ after removing the misfitting item. With 2-item scales, Spearman-Brown's ρ is more appropriate than Cronbach's α (Eisinga, Grotenhuis, & Pelzer, 2012).

Morris (2010) reviewed and empirically validated a wide range of instruments for the sources of teaching self-efficacy. Using item-level information from Morris, I selected 6 to 8 of the best performing items to address each source, adapting the wording in each to focus on multiplicative reasoning rather than teaching in general. These revised items comprised the TSE

Sources instrument. For example, the vicarious experience question "I often try to visualize myself working through the most difficult teaching situations" became "When I am preparing to teach topics involving fractions, ratios, and proportions, I often try to visualize myself working through the most difficult teaching situations." (Recall that vicarious experience can involve recollection of a past performance as well as observation of another's performance.) Sample items for each source instrument are provided in Table 4. A second question in the validation work described below was whether TSE Sources would predict TSE Beliefs in the ways stipulated by social cognitive theory.

Other survey items and instruments. The TTMR survey and TX-TEDS studies included a range of items and instruments on the background, preparation, and school context of participants. Both demographic and certification data were provided by the Texas Department of Education and reflect official records. These variables include *age*, *ethnicity, gender*, and *certification grade level*. Age was reported in years; the categories of the ethnicity variable follow the corresponding U.S. Census question; and the gender and certification grade level variables were included in the analysis using binary dummy indicators (i.e., values of 1 or 0 for each category based on whether that individual was included). There were six certification grade level categories in the original data: EC–4 (n = 201), EC–6 (n = 6), 4–8 (n = 245), EC–8 (n = 15), 8–12 (n = 196), and 4–12 (n = 16). Three certification grade level categories (EC–6, EC–8, & 4–12) were too small for independent analysis, so I combined each of these with the most similar remaining category, resulting in an elementary category (EC–4 and EC–6), a middle grades category (4–8 and EC–8), and a high school category (8–12 and 4–12). Data for all four of these variables were obtained from the Texas Department of Education and reflect official records.

Participants in both TX-TEDS surveys reported the institution where they obtained their bachelor's degree, and I used rating data describing the selectivity of participants' undergraduate institution from a 6-point selectivity rating scale (Barron's Educational Services, 2001). The scale ranged from 1 (*noncompetitive*) to 6 (*most competitive*), and was calculated for each school by Barron's Educational Services from various factors including students' average SAT and ACT scores, minimum class rank of accepted students, percentage of incoming students in the top 40% of their high school class, and percentage of applicants accepted. Educational economists have used *undergraduate selectivity* as a proxy for cognitive ability (see, e.g., Humphrey et al., 2008); it may also reflect students' opportunities for education. In the following analysis, this variable was used to characterize the analytic sample and as an auxiliary variable to mitigate the uncertainty due to missing data; both uses were warranted under either interpretation. Participants were asked on the TTMR survey whether they had begun teaching full time before receiving full certification, and a dummy variable was used to incorporate this *early entry* variable into the analyses.

An important piece of teachers' preparation and certification program is their academic preparation during university coursework. I created a composite *perceived academic preparation* variable to summarize teachers' responses to the following question on the TX-TEDS surveys, "How well prepared academically do you feel you are—you feel you have the necessary disciplinary coursework and understanding—to teach each of the following at the grade level you are currently teaching?" To correspond with the focus on multiplicative reasoning in the outcome instruments, I selected a subset of the available items for the composite academic preparation variable. Teachers in the primary grades (EC–6) were asked about their academic preparation for teaching (a) fractions and (b) decimal topics, among others. Teachers

in the secondary (4–12) grades were asked about their academic preparation for teaching (a) fraction and decimal topics and (b) proportion and ratio topics, among others. The teachers responded to each item on a 4-point Likert scale (1: not well prepared; 2: somewhat prepared; 3: fairly well prepared; and 4: very well prepared). To create the composite variable, I took the average of items *a* and *b* described above or the score of a single item (*a* or *b*) if one was missing.

On the one hand, the composite perceived academic preparation variable has different absolute meanings across different grade levels: The score reflects elementary teachers' reports on fraction and decimal topics separately but secondary teachers' reports on these topics together and then on the additional topics of proportion and ratio. On the other hand, the composite variable has a clear meaning for the analysis that is coherent across certification grade levels—the variable reflects the sense of academic preparedness teachers reported relative to the multiplicative reasoning topics they were most likely to be asked to teach at the grade level for which they were certified. The perceived academic preparation variable was entered into the analyses as a continuous variable.

The teachers were also asked about their student teaching experiences. Two variables related to the practicum were used in the analyses. First, the participants were asked about the frequency of their experience with three high quality student teaching experiences: enrollment in coursework connected to the practicum, meeting individually with mentor during practicum, and observing fellow student teachers and discussing their practice. These items were on a 3-point scale (1: never; 2: sometimes; 3: always) and were used as indicators for a variable called *high quality experiences* and modeled as a latent factor ($\alpha = .89$). Second, the participants reported the student teaching *length* they experienced in weeks, "How many weeks did you spend in your

teacher preparation program actually teaching as part of a practicum for more than one day a week?" The distribution of reported student teaching length is shown in Figure 2.

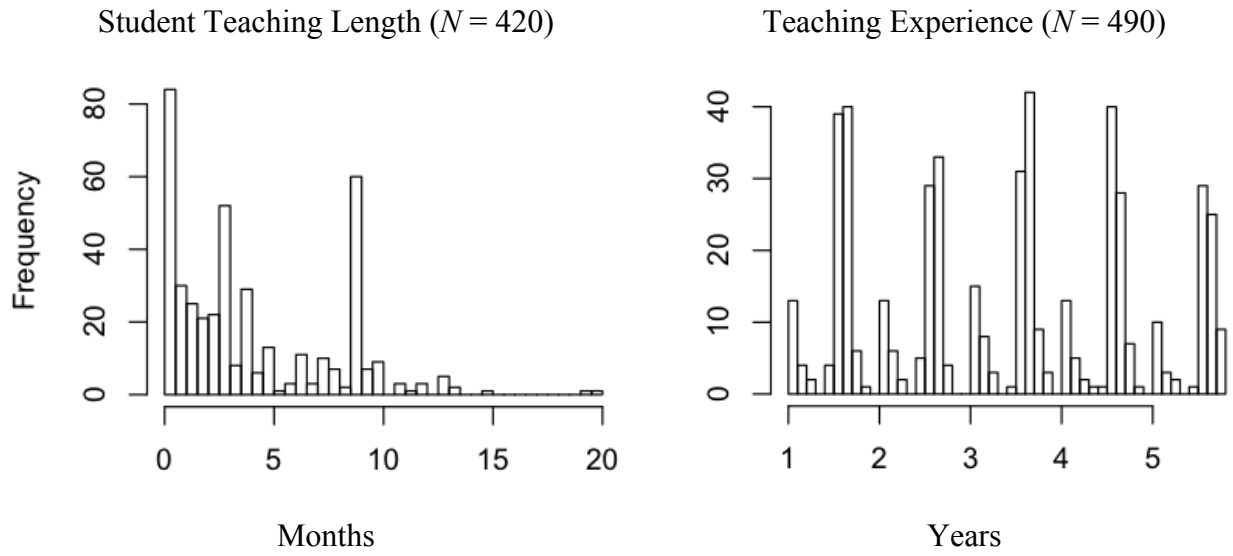


Figure 2. Histograms of the independent variables for student teaching length and teaching experience.

The teachers were asked several questions about their experience teaching. First, the variable *teaching experience* was calculated from official records of initial certification and the date each participant took the survey (administration dates ranged from July 2011 to April 2012). Adding the survey administration lag time more accurately represented differences in teaching experience between participants in the same certification cohort, and the ten spikes in the histogram correspond to members of the five certification cohorts who participated in either TX-TEDS-1 (July to October 2011) or TX-TEDS-2 (November 2011 to May 2012) studies (see Figure 2).

The second variable describing teachers' experience teaching had to do with the content they were teaching. The TX-TEDS survey included a question on the number of class periods in the 2010–2011 school year that teachers had spent teaching various topics. These data were combined to form a variable called *topical experience*. Teachers in the primary grades were

asked how often they had taught (a) fractions and (b) decimal topics, among others. Teachers in the secondary grades were asked how often they had taught (a) fractions and decimals and (b) proportions and ratios, among others. Responses to each item were entered on a 6-point scale (1: 0 periods; 2: less than 1 or 1 period; 3: 2 to 5 periods; 4: 6 to 10 periods; 5: 11 to 15 periods; 6: 15 or more periods).

I summed the codes from Topics *a* and *b* described above to combine these data into a single measure of topical experience across grade levels, but this sum was skewed by the fact that 1 point on the original scale meant an increase of less than a class period of experience at the low end of the scale, but also meant an increase of 4 or 5 class periods at the high end of the scale. To mitigate the shift in meaning for a 1-point difference across the scale, I subtracted 1 from each sum for each item with a code of 2 or greater. For example, a teacher with a code of 1 (0 periods) and 3 (2 to 5 periods) would have a combined code of 4 – 1 or 3 (2 to 5 periods), and a teacher coded 4 (6 to 10 periods) on both items would have a combined code of 8 – 2 or 6 (15 or more periods). The first six codes of the composite variable obtained in this way had approximately the same meaning as the original six codes. I defined additional codes for composites codes of 7 and above (7: 17 or more periods, 8: 21 or more periods, 9: 26 or more periods, and 10: 30 or more periods). These definitions necessarily overlap because the original codes indicate a range of possible values. Such codes were necessary to reflect the probable differences in frequency of topical experience of participants who had reported codes above 3 on the original scale for both topics in the composite. Using a composite code for the topical experience variable across grade levels even though different teachers were asked about different topics by grade level is defensible because opportunities to teach multiplicative reasoning topics

at each grade level were likely constrained to the topics each teacher was asked about. This composite code was entered into the analyses as a continuous variable.

The third variable describing each teacher's experience teaching was *collaborative activity*. This variable was included in the analysis to reflect differences in teachers' professional work environment. This variable was modeled in analyses as a latent factor with five indicators ($\alpha = .78$), each a 5-point scale (1: never; 2: once a year; 3: once or twice each semester; 4: once a month; 5: once each week). These five indicators asked about the frequency of professional activities that might support teachers' learning with colleagues, especially activities focused on student thinking. Teachers were asked how frequently in the past 3 years they had done each of the following activities with colleagues: analyzed sample student work, sought advice about instructional issues, discussed teaching practice, discussed the strengths or needs of specific students, and discussed student assessment data to make instructional decisions.

Distributional statistics for the independent variables used in the analysis are shown in Table 5. The variables are grouped by analytic purpose and meaning. The first group (age and undergraduate selectivity) were used as auxiliary variables to mitigate bias from missing data. The other three groups (preparation, student teaching, and teaching) describe categories of teachers' experience that address the second research question of the study. The analytic methods I used assume that all independent variables have normal distributions, and the low skewness and kurtosis values for these variables (see Table 5) indicated that all of these variables were approximately univariate normal.

Table 5

Auxiliary and Independent Variables Used in Multivariate Regression Analyses

Variable	*M*	*SD*	Skewness	Kurtosis
Auxiliary				
Age (years)	39.37	10.97	0.65	-0.66
Undergraduate selectivity	3.17	1.36	0.06	-0.80
Preparation				
Perceived academic preparation	3.60	0.70	-1.78	2.55
Student teaching				
Length (months)	4.14	3.76	0.90	0.34
High quality experiences (α = .89, ρ = .68) [a]				
1	2.12	0.84	-0.24	-1.54
2	2.28	0.69	-0.44	-0.88
3	1.84	0.74	0.26	-1.15
Teaching				
Topical experience	6.0	3.2	-0.12	-1.36
Teaching experience (years)	3.4	1.4	0.07	-1.19
Collaborative activity (α = .78)				
1	2.67	1.34	0.39	-1.01
2	4.04	1.14	-0.99	0.00
3	3.91	1.14	-0.90	0.09
5	3.69	1.30	-0.74	-0.56
8	3.31	1.13	-0.49	-0.38

[a] Item 1 was not used in PQ composite because of differential additive effects across the early entry status grouping; see the Measurement Invariance section for more details. For 2-item scales, the Spearman-Brown reliability coefficient ρ is more appropriate than Cronbach's α (Eisinga et al., 2012).

Characteristics of the Analytic Sample

The analytic sample for this study includes all responders (N = 492) to the TTMR survey, and thus, the analytic sample combined volunteers from among the participants of both TX-TEDS studies. Combining the TTMR respondents from both TX-TEDS studies increases the precision of the analytic results of this study because of the larger N, but it also means that the analytic sample and findings based on this sample are not representative of the population of K–12 mathematics teachers in Texas. I chose to combine these groups because the primary purpose

of the study was to evaluate the validity of measures and to describe possible relationships between teachers' experience and their mathematical proficiency for teaching to inform further research. The primary concern was to obtain a sample that included a large portion of the variation in the population of Grades K–12 teachers certified in the last 5 years in Texas, and this goal was almost certainly achieved with the data available in the combined analytic sample. The characteristics and composition of the analytic sample and the samples for each TX-TEDS study are described in detail in Table 6.

The sampling strategy of the TX-TEDS1 study was a census, but the sample is better understood as a convenience sample because the response rate (6%) was extremely low. The sampling strategy for TX-TEDS2 was stratified sample randomized by grade-level strata, and thus is more likely to be representative (within grade-level) of the population of Texas mathematics teachers. The low response rate of 23% for that study, however, precluded high confidence of accurately generalizing results to the population. The demographic variables in Table 6 provide further evidence that the two TX-TEDS studies did not sample the same population. The biggest difference was in the distribution of participants by certification grade level. Almost three-quarters of the TX-TEDS-1 participants were elementary certified whereas less than one-quarter of the TX-TEDS-2 participants were certified for those grades. This difference may underlie other apparent differences. For example, the TX-TEDS1 study had 90 female participants, but TX-TEDS2 had just 64 female participants, which is not surprising given that elementary teachers are more frequently women than men. The TX-TEDS1 study participants were also slightly younger, slightly less Asian, and slightly more African American and Hispanic than participants of the TX-TEDS2 study.

Table 6

Demographic and Certification Characteristics as a Percentage of Two TX-TEDS Study Samples and the Analytic Sample

Characteristic	TX-TEDS 1 (Jul–Oct 2010)			TX-TEDS 2 (Nov–May 2011)			Analytic sample
	Full TX-TEDS1 sample (*n* = 1015)	TTMR volunteers (*n* = 166)	TTMR responders (*n* = 106)	Full TX-TEDS2 sample (*n* = 1937)	TTMR volunteers (*n* = 516)	TTMR responders (*n* = 386)	All TTMR responders (*n* = 492)
Gender - female	90	90	90	64	68	68	73
Ethnicity							
African American	9	6	6	8	8	7	7
Asian	2	2	3	6	3	2	2
Hispanic	19	17	16	18	15	15	15
Native American	1	2	1	<1	<1	<1	<1
Other	2	1	2	2	1	1	1
White	67	73	73	67	73	75	74
Certification type							
Alternative	29	29	32	57	57	57	52
Out of state	9	12	11	19	20	22	19
Standard	62	59	58	24	23	21	29
Certification year							
2006	21	17	14	18	17	17	16
2007	20	26	22	18	18	19	20
2008	21	20	24	21	22	22	23
2009	20	18	20	20	20	19	19
2010	19	20	21	23	24	23	22
Certification grade							
EC–4 & EC–6	74	66	63	22	19	19	28
Grades 4–8 & EC–8	20	23	25	45	43	42	38
Grades 8–12 & 4–12	6	11	12	33	38	40	34

Two important questions about the analytic sample are relevant for the subsequent analyses and for understanding the implications of the results: How similar were the TTMR volunteers to the participants in the TX-TEDS studies? How similar were the TTMR responders (the analytic sample) to the nonresponding volunteers? These questions address the possibility of two forms of selection bias. First, I was concerned that those volunteering to take the TTMR survey differed systematically from those who were invited. Second, I was concerned that TTMR respondents differed systematically from the TTMR volunteers (e.g., perhaps low-knowledge volunteers chose not to take the survey more often than high-knowledge volunteers after seeing the MKT items). The available data showed that neither of these concerns was founded; overall, the TTMR volunteers were similar to the respective TX-TEDS study participants who might have volunteered instead, and those who chose to respond to the TTMR survey were very similar to the TTMR volunteers who did not respond.

From Table 6, it is evident that the TTMR volunteers from each TX-TEDS study had similar demographic characteristics to the corresponding TX-TEDS study sample except in mean age: Each group of TTMR volunteers tended to be slightly older on average than the corresponding TX-TEDS study participants. There were no substantial differences between each group of TTMR volunteers and the corresponding TX-TEDS study sample with respect to certification variables, or with respect to age or undergraduate selectivity (see Table 7).

To address the degree of balance between TTMR volunteers and the eventual responders (the analytic sample), I used the MatchBalance function of the R package Matching that is designed to compare treatment and control groups for equivalence on multiple variables using *t*-tests, bootstrap Kolmogorov-Smirnov tests, and summary statistics for the empirical cumulative distribution function, and an empirical quantile-quantile plot of categorical and continuous

variables. There were no significant differences on any of the variables between TX-TEDS1 volunteers who responded to the TTMR survey and those who did not. There were two variables that exhibited significant differences between the TX-TEDS2 volunteers who responded to the TTMR survey and those who did not. First, relatively more Grades 4–8 and EC–8 teachers responded to the TTMR survey than those who did not (38% versus 26%). Second, relatively more teachers certified in 2007 responded to the TTMR survey than those who did not (20% versus 12%). Finally, I compared all teachers in the analytic sample with all the volunteers who did not respond to the TTMR survey. That analysis revealed three significant differences: Teachers in the analytic sample were relatively older than nonresponding volunteers (with mean age 39 years versus 37), relatively less likely to be elementary Grades EC–4 or EC–6 certified (28% versus 36%), and relatively more likely to be secondary Grades 4–12 or Grades 8–12 certified (34% versus 24%).

Table 7

Age and Undergraduate Selectivity of Participants by TX-TEDS Study and Analytic Sample

Characteristic	TX-TEDS 1 (Jun–Oct 2010)		TX-TEDS 2 (Nov–May 2011)		Analytic Sample
	TTMR volunteers *M* (*SD*)	TTMR responders *M* (*SD*)	TTMR Volunteers *M* (*SD*)	TTMR responders *M* (*SD*)	All TTMR responders *M* (*SD*)
Age (years)	41 (12)	41 (12)	38 (11)	39 (11)	39 (11)
Undergraduate selectivity (1: noncompetitive to 6: most competitive)	2.9 (1.3)	2.9 (1.3)	3.2 (1.3)	3.3 (1.4)	3.2 (1.4)

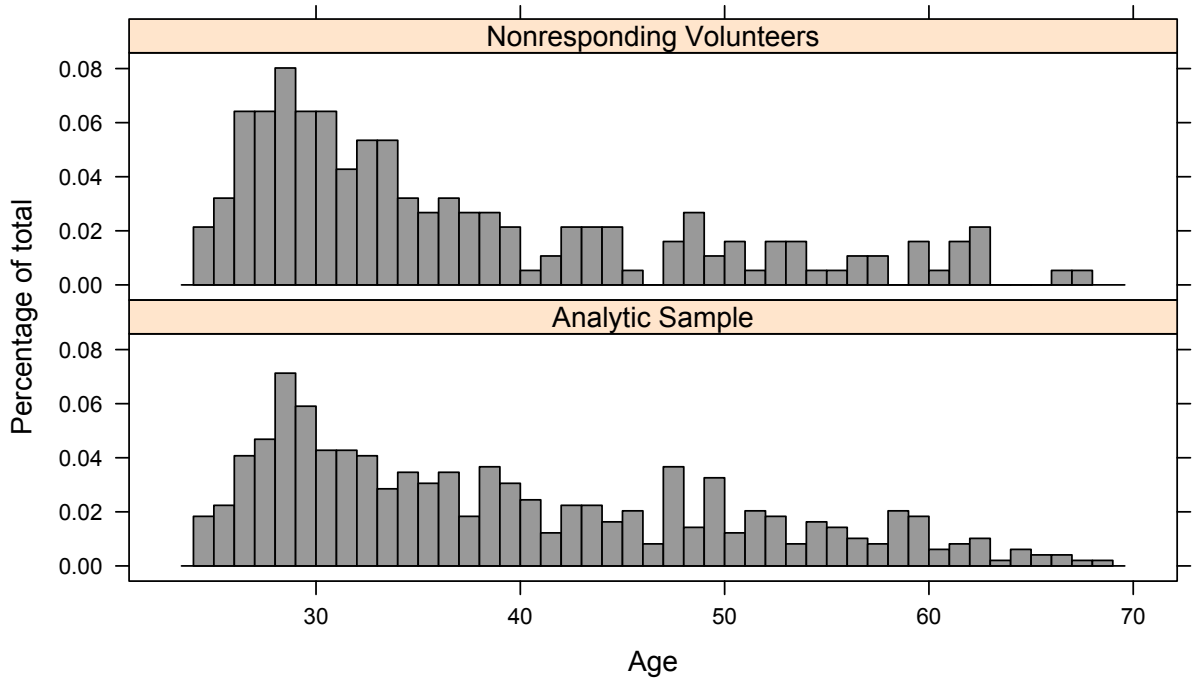


Figure 3. Histograms of participant age for nonresponding volunteers and the analytic sample.

I explored the differences in participation by age and certification grade level by examining histograms and summary statistics for each group to gain a better understanding of how the analytic sample might have differed from the group of nonresponding volunteers. Histograms of participants' age in the analytic sample and among nonresponding volunteers show that volunteers younger than 30 were less common in the analytic sample than among nonresponding volunteers. In addition, volunteers older than 40 were more common in the analytic sample than among nonresponding volunteers (see Figure 3). Next, I compared the analytic sample and nonresponding volunteers on the variable from the TX-TEDS surveys with the most relevance for the planned analysis: their reported level of preparedness to teach fractions (see Table 8). This variable provided no evidence that the differential response pattern at each certification grade level was related to the teachers' sense of preparedness to teach fractions.

Table 8

Preparedness to Teach Fractions by Certification Grade Level and Participant Subgroup

Subgroup	Mean preparedness to teach fractions (*SD*)	
	Analytic sample	Nonresponding volunteers
Elementary generalists certified Grades EC–6 (n = 6) & Grades EC–4 (*n* = 201)	3.28 (0.80)	3.54 (0.69)
Middle grades math specialists certified Grades EC–8 (*n* = 15) & Grades 4–8 (*n* = 245)	3.53 (0.81)	3.65 (0.57)
Secondary math specialists certified Grades 4–12 (*n* = 16) & Grades 8–12 (*n* = 196)	3.85 (0.45)	3.77 (0.64)

In summary, neither of the two sub-samples from the TX-TEDS studies nor the combined analytic sample were representative of the population of Grades K–12 mathematics teachers in Texas that were certified in the last 5 years. There were no differences between TX-TEDS study participants who volunteered to take the TTMR survey and those who did not, suggesting that there was no additional selection bias associated with volunteering to take the TTMR survey. There were some small differences between TTMR volunteers and responders. Those who responded to the TTMR survey were slightly older and were relatively less likely to be elementary certified and relatively more likely to be middle or secondary certified. On the other hand, the TTMR responders did not on average have more experience or a greater sense of preparation for teaching fractions than the nonresponding volunteers. Together these findings support increased confidence that the analytic sample reflected a range of school and preparation experiences of the population of Grades K–12 teachers certified in the last 5 years in Texas and that the selection bias affecting the analytic sample was generally limited to the selection bias associated with the sampling for the TX-TEDS studies.

Missing Data

Because of missing data, the size of the analytic sample available for any given analysis did not attain the total of 492. Omitted responses were grouped toward the end of the survey, which suggested that almost all of the omissions occurred because participants stopped before reaching the end of the survey. To estimate the IRT parameters for the MKT instrument, I used data from the 409 respondents that had answered at least one item on that instrument. For the TSE instruments, the data did not satisfy the normality assumptions, and I used estimation methods that are robust to violations of the normality assumptions but require complete responses. There were 426 respondents with complete responses available on the TSE Beliefs instrument. The TSE Sources instruments were given only to volunteers from the TX-TEDS2 study, so the available sample for this instrument was limited to 320. To perform the final SEM analyses, I used the software program MPLUS and full information maximum likelihood estimation with robust standard errors, which allows all respondents (even those with missing data) to be incorporated in the analysis.

Results

I describe the results of the study in this section, addressing the each research question in turn. First, I describe the psychometric work and validity argument supporting the MKT instrument and the TSE Beliefs instrument. Next, I describe the variation in mathematical proficiency for teaching multiplicative reasoning in the analytic sample in relation to participants' background, preparation, and school experience.

Instrument Reliability and Validity

One major purpose of this study was to establish the reliability and validity of domain-specific instruments of mathematical proficiency for teaching. The first research question for this study was as follows:

1. How valid and reliable are the instruments of content knowledge for teaching and teaching self-efficacy that were adapted to target the domain of multiplicative reasoning?

Validity has been discussed extensively in the psychometric literature and theory and practice are evolving (e.g., Messick, 1988; Kane, 2001, 2004). Following Kane (2001, 2004) and Schilling and Hill (2007), I consider the elemental and structural validity of both the MKT and the TSE Beliefs instruments. Elemental validity concerns valid measurement at the item level. For the MKT instrument, for example, this meant finding evidence that the individual items actually measured teachers' mathematical knowledge for teaching in the domain of multiplicative reasoning, and moreover that how teachers answered an item (i.e., the reasoning that they used) was aligned with the choice they selected and what that choice was supposed to signify via the official item key (i.e., possessing or not possessing the MKT in question). Structural validity concerns valid measurement at the instrument level. The major questions of structural validity for this study concerned whether the MKT instrument had a single unidimensional structure (as expected) and whether the TSE Beliefs instrument had the two-factor structure predicted by theory.

A third kind of validity discussed by Schilling and Hill (2007) was ecological validity. Ecological validity is similar to construct validity and concerns the relationship of an instrument with other instruments and with other sources of empirical data. I tested the ecological validity of

the TSE Beliefs instrument by examining its relationship to the TSE Sources instrument. Assessing the structural validity of the MKT instrument was outside the scope of this study.

MKT instrument. The validity argument for the MKT instrument is based on my analysis of the content the items included in the instrument (see Table 3) and the item development work conducted by the Measures of Effective Teaching project (Bill & Melinda Gates Foundation, 2010). The MKT item development work was overseen by the Educational Testing Service. The rigorous process led from item writing, to item revisions or rejection based on expert content review, to item response interviews that elicited teachers' reasoning about the item and their rationale for choosing answers, and concluded with standard psychometric screening based on pilot data. The source items for those items adapted from the DTMR project were subject to similar item-response interviews but had not been psychometrically validated with pilot data. By selecting items for the multiplicative reasoning MKT instrument that were already deemed appropriate for measuring elementary or middle grades MKT, I addressed the key question in this study with respect to the elemental validity of the MKT instrument: Do the items individually provided valid measurement of MKT related to multiplicative reasoning topics? My analysis of the content of these items and review of item-response interview data led me to believe they do tap MKT for multiplicative reasoning. This judgment was informed by the review of literature on students' and teachers' multiplicative reasoning that is summarized in Chapter 2.

The major question with respect to the structural validity of the MKT instrument was whether the subset I had identified was psychometrically coherent. I relied on psychometric evidence to determine that the MKT instrument was unidimensional and that each item was related to the others. In response to initial IRT analyses, I removed 1 of the 26 items from the

instrument (see Appendix B). The final, 25-item, instrument had high internal consistency (Cronbach's alpha of .95), and the final item parameters of the retained items were in an acceptable range (Baker, 2001; see Appendix B for more detail). All items had point-biserial correlations greater than or equal to .2; with a sample size of approximately 400, these values are much higher than the recommended cutoff for inclusion of two standard deviations above 0 (Crocker & Algina, 2006). The estimated abilities of respondents on the MKT instrument were distributed normally (see Figure 4), with more test information and hence smaller errors of estimate for the lower half of the observed score distribution than for the upper half.

TSE Beliefs Instrument. The elemental validity of the TSE Beliefs argument is based on the elemental validity of the instruments from which it was adapted. Enochs and Riggs (1990) and Roberts and Henson (2000) described the item construction and psychometric validation work conducted to produce the items that were modified to create the TSE Beliefs instrument. These items have also been similarly modified for use with preservice mathematics teachers and validated in that context (Enochs, Smith, & Huinker, 2000). In related work, Newton (2009) modified mathematics anxiety and motivation questionnaire items to make them specific to fractions by replacing the word *mathematics* with the word *fractions*. This prior work taken together provides confidence that the modifications I made to adapt the TSE Beliefs items to make them domain-specific have not affected the validity of individual items with respect to construct of teaching self-efficacy beliefs.

The types of validity are not independent, and lack of validity at the elemental level would likely cause a lack of evidence for structural and ecological validity. To examine the structural validity of the TSE Beliefs instrument, I used confirmatory factor analysis and found the expected two-factor model showed no evidence of misfit. It demonstrated a statistically

significant improvement in fit over a one-factor model of TSE Beliefs. Similar psychometric work provided strong evidence of the structural validity of the TSE Sources instrument. More details about these results are available in Appendix B.

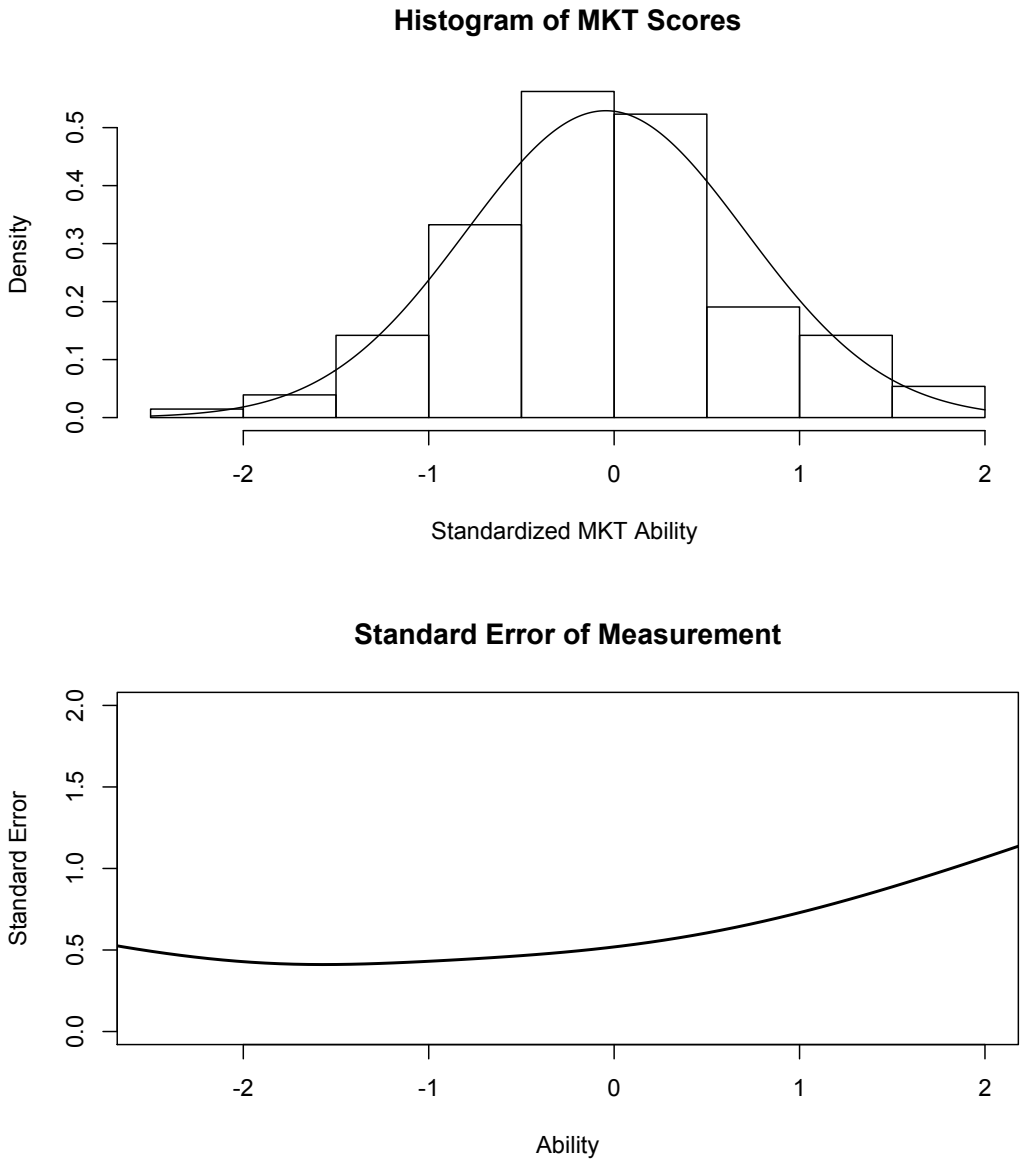


Figure 4. The score distribution and standard error of estimate and over the range of observed ability (θ) for the MKT instrument.

Evidence of ecological validity for the TSE Beliefs instrument came from two findings. First, I expected significant correlations between the factors of the TSE Beliefs instrument (personal efficacy (PE) and knowledge efficacy (KE)), and the four factors of TSE Sources instrument (mastery experience (ME), vicarious experience (VE), social persuasion (SP), emotional and physiological states (EP)). Second, I expected that the factors of TSE Sources could be used to predict TSE Beliefs and furthermore, that they would explain a large amount of the variance in PE and KE. Table 9 shows the correlations between the factors of TSE Beliefs and TSE Sources instruments; they are all statistically significant at the $p = .05$ level except the correlation between vicarious experience and PE. I also used structural equation modeling to regress PE and KE on the TSE Sources, and found that the TSE Sources explained 82% of the variance of PE and 69% of the variance of KE.

Table 9

Correlations Among the Factors of TSE Beliefs and TSE Sources (Raw Scores)

Factor	Knowledge efficacy	Mastery experience	Social persuasion	Vicarious experience	Emotional & physiological states
Personal efficacy	.67*	.73*	.45*	.10	.68*
Knowledge efficacy		.69*	.59*	.23*	.58*
Mastery experience			.56*	.14*	.67*
Social persuasion				.28*	.42*
Vicarious experience					.10*

* $p < .05$.

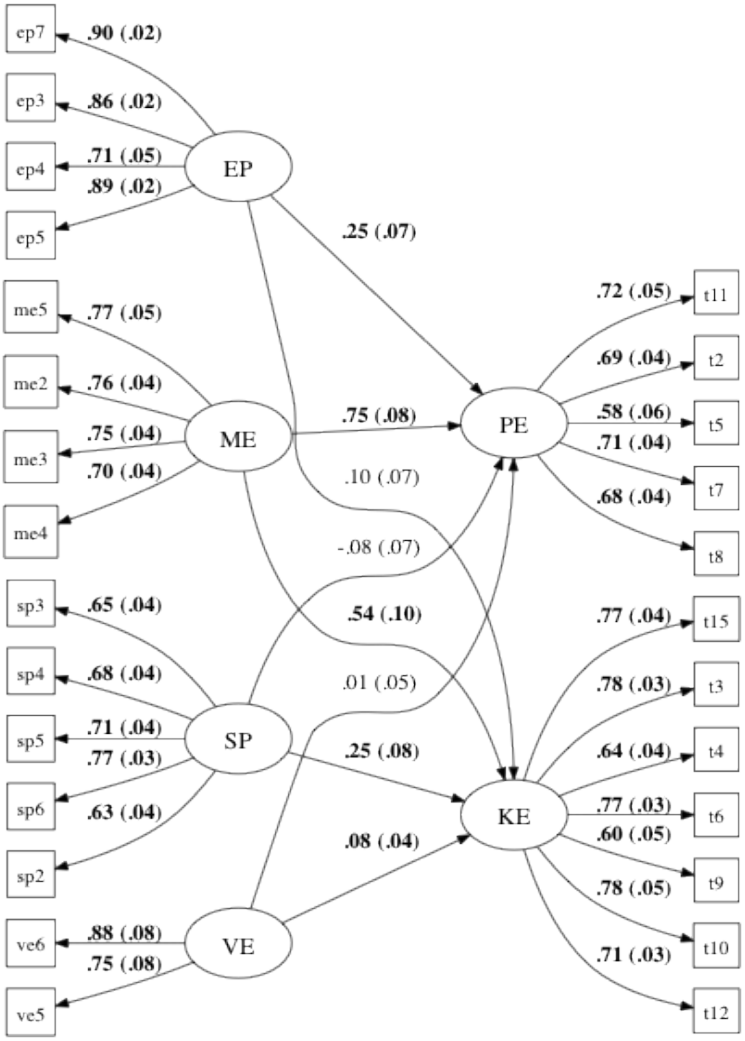


Figure 5. Structural equation model for the factors of TSE Beliefs regressed on the TSE Sources factors. All bold path coefficients were statistically significant at $p < .01$ except KE on VE which was close to statistical significance at the $p < .05$ level ($p = .06$, not shown).

Figure 5 shows the fitted model. Each factor was modeled as a latent (unobserved) variable and is represented by an oval; each item was modeled as a manifest (observed) indicator and is represented by a square. Arrows indicate the predictive paths between these latent and manifest variables. Path coefficients can be read as standardized regression coefficients; for every unit increase in the variable at the tail of a path, one can expect an increase in the variable at the head of the path of magnitude equal to the number of units represented by the path

coefficient. Standard errors are given in parentheses in the diagram in Figure 5, and *p* values can be calculated for each coefficient by dividing the coefficient by the standard error to obtain the appropriate *z*-score.

I used the Satorra–Bentler (SB) χ^2 test statistic to test model fit, which is appropriate when data are nonnormally distributed (Bentler, 2005), and the test of model fit was significant ($\chi^2_{SB} = 409.7$, $df = 309$, $p < .001$). Significant χ^2 tests of model fit indicate that the data do not fit the model perfectly, but this test is overly sensitive, and in practice fit indices are used to evaluate lack of model fit instead (Kline, 2005; Bandalos & Finney, 2010). These indices do not prove that the model fits the data but can indicate whether the model exhibits various kinds of misfit. Kline (2005) recommended that if the model fits, then the comparative fit index (CFI) should be above .90; the upper bound of the root mean square error of approximation (RMSEA) 90% confidence interval should be less than .10; and the square root mean residual (SRMR) should be less than .08. The structural equation model of TSE Beliefs regressed on TSE Sources showed no signs of misfit (CFI = .94, RMSEA = .032 with a 90% confidence interval of .020 to .040, SRMR = .050).

Bandura (1997) theorized that mastery experience is the most powerful source of self-efficacy and that the three other sources influence self-efficacy to a lesser degree. He also argued that individuals would give different emphasis to the different sources of self-efficacy in different contexts. Both of these aspects of social cognitive theory were well reflected in the results of a structural equation model of the TSE Beliefs factors regressed on the TSE Sources factors. First, mastery experience (ME) was the strongest predictor for both factors of TSE Beliefs. Second, the regression (path) coefficients for each source were different for different factors of TSE Beliefs. For example, the emotional and physiological states (EP) factor was a

significant predictor of PE but not a significant predictor of KE. Conversely, social persuasion (SP) is a significant predictor of KE but not a significant predictor of PE. Vicarious experience (VE) predicted KE with a path coefficient of .08 and a standard error of .04 was very close to statistical significance at the .05 level ($p = .06$) but VE did not predict PE with statistical significance ($B = .01$, $SE = .05$, $p = .76$).

Predictors of Mathematical Proficiency for Teaching

In this final section, I report on results pertaining to the second research question:

2. How does mathematical proficiency for teaching multiplicative reasoning topics vary (a) with perceived academic preparation, student teaching, and teaching experience and (b) by early entry status and across certification grade level?

I hypothesized that perceived academic preparation, student teaching, and teaching experience would all have significant positive relationships with mathematical proficiency for teaching.

A central concern in addressing this question was the hypothesized interaction between aspects of certification (early entry and grade level) and the other antecedents of mathematical proficiency for teaching under analysis. I expected no differences in means between the standard entry and the early entry groups, but I did hypothesize that student teaching and teaching experience would be more weakly related to mathematical proficiency for teaching in the group of early entry teachers than in the group of standard entry teachers. With respect to grade level certification, I hypothesized that that perceived academic preparation and student teaching as well as teaching experience might have stronger relationships with mathematical proficiency for teaching multiplicative reasoning in the middle grades group than in the other groups because multiplicative reasoning topics are more prevalent in the middle grades curriculum and thus perhaps more of a focus in middle grades preparation programs and middle grades teachers'

experiences in schools. I also expected differences in the mean mathematical proficiency for teaching multiplicative reasoning between the different certification grade level groups. Past research (e.g., Hill, 2007) and my own experience teaching and conducting interviews with prospective and practicing teachers suggested that secondary certified teachers might have higher MKT and TSE for multiplicative reasoning topics than middle grades teachers, and both might have higher mathematical proficiency for teaching than teachers certified for the elementary grades. At the same time, I was curious as to whether there was evidence that experience teaching in grades with the curriculum focused on multiplicative reasoning topics would moderate initial differences. It seemed plausible to me that a middle grades teacher after 5 years of teaching proportion would have greater mathematical proficiency for teaching multiplicative reasoning topics than a first year, secondary certified teacher assigned to teach statistics and calculus.

Measurement invariance. To make valid comparisons between groups, I needed to ensure that the instruments used to measure the latent traits used in the analysis functioned in similar ways across the two sets of certification groups (grade level and early entry status). I used multiple group CFA analyses to check whether the instruments were invariant across groups. I also used the Mantel-Haenszel (Holland & Thayer, 1988) and the generalized Mantel-Haenszel (Penfield, 2001) chi-square tests to examine differential item functioning in the MKT instrument. (Multiple sample CFA methods require continuous variables—and the rating scale responses from the beliefs instruments were sufficiently continuous for these methods—but the MKT items were dichotomously scored.)

***Polytomously scored instruments*.** Measurement invariance was tested with a sequence of nested CFA models: First, the measurement model was fit with item loadings and item

intercepts for each latent factor allowed to vary by group (Model 1); next, the item loadings were constrained to be equal between groups (Model 2); finally, the item intercepts were constrained to be equal between groups (Model 3). Subsequent models in this sequence have fewer parameters than previous models, so they do not fit the data as well. A chi-square test of model fit was used to test the null hypothesis that the earlier model with more parameters has a significantly better fit.

Tests that fail to reject the hypothesis that Model 1 fits better than Model 2 are evidence of *metric invariance*, and tests that fail to reject the hypothesis that Model 2 fits better than Model 3 are evidence of *strong factorial invariance*. Metric invariance supports valid comparisons of structural path coefficients between groups, and strong factorial invariance supports valid comparisons of group factor means (Meredith, 1993). An intermediate case, where some (preferably most) item intercepts are constrained to be equal across groups, is termed *partial measurement invariance* and is frequently accepted in practice (Byrne, Shavelson, & Muthen, 1989).

Results of the measurement invariance analysis are reported in Table 10 (early entry status) and Table 12 (certification grade level). Both sets of analyses used the same overall measurement model with four correlated factors: Personal Efficacy (PE), Knowledge Efficacy (KE), Quality of Student Teaching (QU), and Collaborative Activity (CA) (see Figure 6). In both tables, Models 1, 2, and 3 refer to the increasingly constrained nested models: all parameters free, item loadings constrained between groups, and item loadings and intercepts constrained between groups, respectively. In both multiple group analyses, the metric invariance assumption held, but the strong factorial invariance assumption (required for testing differences in means) did not hold. It was necessary to relax that assumption and allow some item intercepts

to vary with groups (for both grade level and early entry groupings) to obtain partial factorial invariance models (Model 4) that did not have significantly worse fit than the metric invariance models.

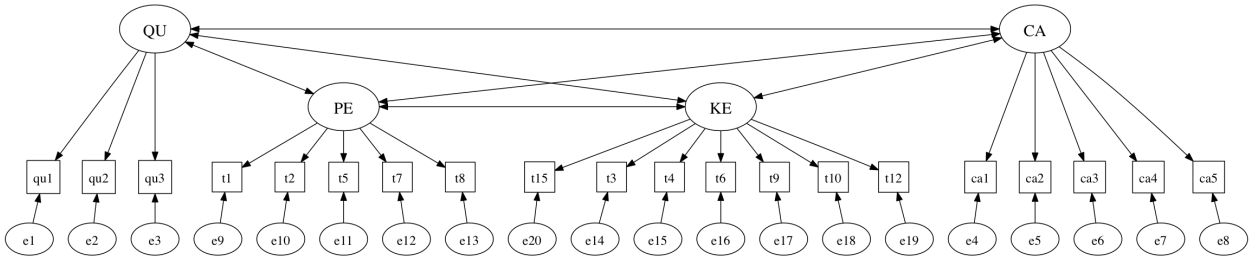


Figure 6. The common measurement model used to test invariance across certification grade level and early entry status groupings; each latent factor is allowed to correlate with all others.

A chi-square test of strong factorial invariance for the early entry grouping determined that there was significant variation (see Table 13), evidence that one or more items had different thresholds across groups after taking into account differences in group means on the corresponding factor. I examined the estimated item intercepts and found that the misfit in this case was caused by Quality of Practicum (QP) Item 1 ("I was enrolled in a course that was connected to my practicum/student teaching."). Early entry teachers often continue to take university courses towards full certification after they have begun teaching fulltime, which is a plausible reason for the observed differential additive effects (see Table 14). Model 4 in Table 13 is a partial measurement invariance model where the intercept for QP Item 1 is allowed to vary by group. The nonsignificant chi-square difference test for Model 4 and Model 2 is evidence that partial measurement invariance held. To prevent the bias that can arise from differential additive effects of items across groups, I did not include QP Item 1 in the composite score for QP in the remaining analyses. Model 4 for the early entry grouping had a significant

chi-square statistic of model fit ($\chi^2_{S\text{-}B}$= 587.265, *df* = 359, *p* = .000), and the other fit indices suggested fit was marginally acceptable (CFI = .90, RMSEA = .06, SRMR = .08).

Table 10

Tests of Multiple Group Measurement Invariance Across Early Entry Status for the Polytomously Scored Instruments: Personal Efficacy (PE), Knowledge Efficacy (KE), Quality of Practicum (QP), and Collaborative Activity (CA)

Model	Equality constraint	χ^2_{SB} [a]	*df*	Comparison	$\Delta\chi^2_{SB}$ test statistic	$\Delta\chi^2_{SB}$ test *df*	*p*
1	None	563.929	328	-	-		-
2	Item loadings	574.513[b]	344	1-2	13.635	16	.635
3	Item intercepts	612.272[c]	360	2-3	38.501	16	.001
4	All intercepts but QP Item 1	587.265[d]	359	2-4	12.470	15	.643

[a] The SB scaling correction factors were as follows: 1.015 (Model 1), 1.026 (Model 2), 1.023 (Model 3), and 1.026 (Model 4).

Table 11

Mean and Mean Differences of Quality of Practicum (QP) Item Responses by Early Entry Status

Status	QP 1[a]	QP 2	QP 3
Early entry	1.71	2.06	1.70
Standard entry	2.51	2.48	2.04
Difference	0.80	0.42	0.34

[a] Item QP 1 exhibited more than double the difference in means between groups than did the other items.

Similarly, the chi-square test of strong factorial invariance for the certification grade level grouping determined that there was significant variation (see Table 12), evidence that one or more items had different thresholds across groups after taking into account differences in group means on the corresponding factor. I examined the estimated item intercepts and found that the misfit in this case was caused by TSE Beliefs Item 8 ("I am not sure I have the necessary skills to

teach every topic involving fractions, ratios, or proportions."), which loaded on the PE factor. It is plausible that the wording "every topic"—unique to this item—led to the observed differential additive effects across certification grade level groups (see Table 16). Model 4 in Table 15 is a partial measurement invariance model in which the intercept for Item 8 is allowed to vary by group. The nonsignificant chi-square difference test for Model 4 and Model 2 is evidence that partial measurement invariance held. To prevent the bias that can arise from differential additive effects, I did not include TSE Beliefs Item 8 from the composite score for PE in the remaining analyses. Model 4 for certification grade level grouping had a significant chi-square test of model fit (S-Bχ^2 = 850.498, *df* = 554, *p* = .000), and the other fit indices suggested fit was marginally acceptable (CFI = .88, RMSEA = .06, SRMR = .09).

Table 12

Tests of Multiple Group Measurement Invariance Across Certification Grade Level for the Polytomously Scored Instruments: Personal Efficacy (PE), Knowledge Efficacy (KE), Quality of Practicum (QP), and Collaborative Activity (CA)

Model	Equality constraint	χ^2_{SB} [a]	*df*	Comparison	$\Delta\chi^2_{SB}$ test statistic	$\Delta\chi^2_{SB}$ test *df*	*p*
1	None	797.028	492	-	-		-
2	Item loadings	810.925	524	1-2	23.046	32	.876
3	Item intercepts	877.765	556	2-3	66.300	32	.000
4	All intercepts but PE Item 8	850.498	554	2-4	39.573	30	.113

[a] The SB scaling correction factors were as follows: 0.995 (Model 1), 1.016 (Model 2), 1.018 (Model 3), and 1.016 (Model 4).

Dichotomously scored MKT instrument. To assess differential item functioning (DIF) of the MKT instrument between the certification grade level groups and between the standard and early entry groups, I used the difR package in R (Magis, Beland, Tuerlinckx, & De Boeck,

2010) to compute the Mantel-Haenszel (Holland & Thayer, 1988) and the generalized Mantel-Haenszel (Penfield, 2001) chi-square tests for each item and across both the grade level groups and the entry timing groups. I used standard entry teachers (n = 190) as the referent group and early entry teachers as the focus group (n = 149) when evaluating DIF between the entry timing groups. Because the MKT items were produced for middle grades teachers, I used the middle grades group (n = 185) as the referent group and the elementary (n = 137) and high school (n = 167) groups as focus groups when evaluating DIF between the certification grade level groups. I found that none of the 25 items exhibited significant DIF for either grouping (see Table 14). This finding is strong evidence that the instrument is behaving in a similar way across groups and that group comparisons using the instrument are meaningful.

Table 13

Mean and Mean Differences of Personal Efficacy (PE) Item Responses by Certification Grade Level

Grade level	TSE 2	TSE 5	TSE 7	TSE 8[a]	TSE 11
Elementary (EC–4 & EC–6)	3.57	3.65	3.17	2.98	3.61
Middle (4–8 & EC–8)	3.89	4.11	3.88	4.04	4.14
High (8–12 & 4–12)	3.96	4.09	4.08	4.39	4.24
Difference middle, elementary	0.32	0.46	0.71	1.06	0.53
Difference high, middle	0.07	-0.02	0.20	0.35	0.10

[a] Item TSE 8 exhibited approximately double the difference in means between groups than did the other items.

Multivariate regression analysis. Structural equation modeling (SEM) can be used to test how well hypothesized models fit data. I hypothesized that the teachers' student teaching and experience would account for a significant amount of the variation in their mathematical proficiency for teaching. To test my hypotheses about the predictors of mathematical proficiency for teaching, I used SEM in which the structural part of the model was a multivariate

regression of the three correlated factors of mathematical proficiency for teaching on preparation and experience variables. Then I used Wald tests to evaluate the hypotheses.

Table 14

The Mantel-Haenszel (MH) and Generalized MH Chi-Square Test for Differential Item Functioning Across Certification Groupings

	Entry timing		Grade level	
MKT item	MH statistic	*p*	Generalized MH statistic	*p*
1	2.367	.124	1.256	.534
2	0.022	.881	0.674	.714
3	0.688	.407	1.397	.498
4	0.025	.874	0.327	.849
5	1.966	.161	0.975	.614
6	0.866	.352	0.580	.748
7	0.008	.927	0.578	.749
8	0.003	.960	0.133	.936
9	1.752	.186	1.598	.450
10	0.009	.925	2.222	.329
11	1.050	.305	0.868	.648
12	0.199	.656	1.618	.445
13	0.216	.642	1.311	.519
14	0.002	.965	0.004	.998
15	0.001	.970	1.010	.604
16	0.176	.675	0.155	.925
17	1.021	.312	0.461	.794
18	1.189	.276	0.947	.623
19	0.001	.979	0.671	.715
21	0.055	.815	0.872	.647
22	0.025	.874	1.613	.447
23	0.331	.565	0.170	.919
24	1.061	.303	0.650	.723
25	0.312	.576	0.450	.799
26	0.446	.504	1.256	.534

Auxiliary variables. The auxiliary variables (age and undergraduate selectivity) and the other independent variables used in this analysis are summarized in Table 4. Auxiliary variables can be used to minimize bias and enhance the efficiency (power) of analyses that include missing data, especially when the variables are correlated with the variables that have missing values or

are correlated with missing data patterns (Enders, 2010). Age was related to missing data patterns, as older participants were more likely to finish the test. Gender was also correlated with the self-efficacy outcomes, and undergraduate selectivity was correlated with the MKT outcome. I decided that age and undergraduate selectivity were best used in the analysis to mitigate the problem of missing data rather than as additional covariates, because all the outcomes were subject to missing data and because the relationships of these variables to the outcome variables did not address the research questions.

Model fit and selection. I evaluated four SEMs to address the research question. First, I needed to ensure that the measurement model (and especially with the revised PE and practicum quality scales) was still exhibiting adequate fit before proceeding with the analysis of the structural portion of the models. The (revised) measurement model is denoted Model 0 in Table 15. Next I fit Model 1, which included a full structural model: The correlated outcomes were regressed on all the predictors in Table 4, and the predictors were allowed to be correlated. The purpose of Model 1 was to evaluate the early entry indicator as a predictor of mathematical proficiency for teaching. Because early entry was not a statistically significant predictor, I then fit Model 2, which was simpler because it did not include the early entry indicator. Model 2 allowed tests of whether teaching experience and practicum length at different grade levels contributed differently to mathematical proficiency for teaching. I did not find a significant interaction between grade level and experience or between grade level and practicum length, so I then fit the last (and most parsimonious) model to obtain estimates of the regression coefficients (Model 3). This model included just 4 independent variables: the grade level indicators, perceived academic preparation, topical experience, and teaching experience.

Table 15 summarizes information describing the fit of the measurement CFA and the three SEMs. Because these models have different predictors, they are not nested and cannot be compared using likelihood ratio tests. Instead, the Bayesian information criterion (BIC) and Akaike's information criterion (AIC) can used to compare the fit of nonnested models. For both criteria, the model in a set of putative models that has the smallest criterion is statistically preferred. Both AIC and BIC indicate that Model 2 is statistically preferred over Model 1 and that Model 3 is preferred over Model 2.

Table 15

Model Fit Indices for the Measurement CFA Model and SEMs of Mathematical Proficiency for Teaching

Model	χ^2_{SB} [a]	*df*	*p*	CFI	RMSEA [b]	SRMR	AIC	BIC
0: Measurement CFA	113.28	43	.000	.948	.058 (.045, .071)	.035	-	-
1: Early entry & grade level SEM	392.85	268	.000	.949	.031 (.024, .037)	.028	40441	42584
2: Grade level SEM	350.47	232	.000	.951	.032 (.025, .039)	.027	33540	35356
3: Final SEM	198.52	97	.000	.949	.046 (.037, .055)	.033	22152	22710

[a] The SB scaling correction factors were as follows: 1.326 (Model 0), 1.037 (Model 1), 1.057 (Model 2), and 1.152 (Model 3).
[b] The 90% confidence interval for the RMSEA fit index is indicated within parentheses.

I also examined the fraction of variance in each factor of mathematical proficiency for teaching explained by the three models (see Table 16). As I expected, the models with more parameters explained a greater amount of the variance in each factor; however, the differences were not great, especially between Model 1 and Model 2. Although these models explained less

of the variance in PE and KE than the model of TSE regressed on the sources of TSE, none of the decreases in R^2 was statistically significant.

Because .20 to .30 of the variance in mathematical proficiency for teaching was accounted for by preparation and experience, the models had substantial explanatory power, especially as the set of covariates in these models did not include pretreatment covariates for the outcome measures such as the TSE or MKT scores of the participants before they began teaching. The variance of MKT explained was also remarkable, even for the most parsimonious model. In a study using multiple regression to model the MKT of elementary school teachers, Hill (2010) reported an adjusted R^2 between .10 and .14.

Table 16

Variance of Mathematical Proficiency for Teaching Explained by Each Model

Model	PE R^2 (*SE*)	KE R^2 (*SE*)	MKT R^2 (*SE*)
1	.298 *** (.052)	.468 *** (.042)	.238 *** (.034)
2	.298 *** (.052)	.469 *** (.042)	.229 *** (.031)
3	.237 *** (.047)	.397 *** (.041)	.194 *** (.032)

*** $p < .001$

Wald tests of hypotheses. The multivariate Wald test statistic describes the change in model chi-square when one or more parameters are constrained (i.e., set equal to zero or to some other value). The statistic has a chi-square distribution with k degrees of freedom, where k is the number of constrained parameters (Kline, 2005, p. 148). The Wald test has a null hypothesis of no loss of fit, and allows a researcher to determine whether the tested parameter constraints significantly decrease the fit of the model to the data. Thus, significant Wald test statistics indicate that the constrained parameters result in a statistically significant loss of model fit and nonsignificant Wald test statistics imply one cannot reject the null hypothesis of no loss of fit.

I used Wald test statistics in conjunction with the three SEMs to evaluate the hypotheses I had about the predictors of mathematical proficiency for teaching. The Wald test statistics and outcomes are summarized in Table 17. For Model 1, I wanted to know if early entry was a significant predictor of any of the factors of mathematical proficiency for teaching. I also wanted to know if the interaction between early entry and teaching experience or student teaching length was different than zero. For example, a significant positive interaction between early entry and teaching experience on the MKT outcome might indicate that the relationship between teaching experience and MKT was stronger for early entry teachers. Instead, I found that the decrease in model fit after constraining all 9 of these parameters to zero was not statistically significant ($p = .239$). This result supported my decision to not include the early entry indicator in subsequent models.

I used Wald tests with Model 2 to evaluate all six predictors of mathematical proficiency: three related to the preparation (and practicum), and three related to teaching experience. I performed two general kinds of tests. First, I used a Wald test to examine the change in model fit after constraining all regression coefficients for a predictor to zero to evaluate whether the predictor contributed significantly to the estimates of mathematical proficiency for teaching. If the test was statistically significant (rejecting the null hypothesis that the constrained parameters did not reduce fit), I next used a Wald test to examine the change in model fit if the coefficients for the parameter were equal for each outcome across all three grade levels. I found that practicum length, practicum quality, and collaborative activity did not improve model fit, whereas preparation, teaching experience length, and topical experience did (see Table 17). The Wald tests for constraining the coefficients for these parameters to have the same value across grade levels, however, led me to conclude that the null hypothesis of similar fit could not be

rejected. This result supported my decision to consider the simpler Model 3, in which only one coefficient for each parameter regressed on each outcome was estimated across all three grade levels.

In Model 3, I used Wald tests to examine whether the middle and high school grade level indicators were equal on the outcome measures (see Table 17). Constraining the coefficients for these indicators to be equal on the TSE factors did not significantly reduce model fit ($p = .173$), but constraining their coefficients to be equal on MKT did ($p = .020$). Finally, I checked whether the predictive contribution of the preparation, teaching experience, and topical experience were different for PE and KE; constraining these coefficients to equality resulted in a significant Wald test ($p = .049$), and I rejected the hypothesis that these predictors contributed equally to the factors of TSE. In the next section I discuss findings related to values of the estimated parameters in Model 3.

Final model parameters. Table 18 summarizes the parameter estimates for Model 3. All parameter estimates were statistically significant at the $\alpha = .01$ level except for the coefficient for teaching experience on MKT and the coefficients for topical experience on the two factors of TSE. The parameter estimates are standardized and thus carry information about the associated effect size. For example, the coefficients for the grade level indicators express mean differences between each grade level and the elementary certified (EC–4 or EC–6) teachers in the sample. These differences were large, about 0.59 *SD* for PE, 0.92 *SD* for KE, and 0.50 *SD* for MKT. Perceived academic preparation had an effect size of medium to large, ranging from 0.20 *SD* for MKT to 0.67 *SD* for KE. By contrast, the only significant regression coefficient for topical experience had a negative sign and a very small effect size (0.03 *SD*). This result may reflect the

likely higher probability that low knowledge high school certified teachers are more often assigned to teach classes that involve multiplicative reasoning topics.

Table 17

Wald Tests of Hypotheses Across the Models Predicting Mathematical Proficiency for Teaching

Path coefficient hypothesis	χ^2_Δ	*df*	*p*
Model 1: Early entry & grade level SEM			
Early entry indicator variable and the interaction terms for early entry with practicum length and teaching experience length on PE, KE, & MKT are all zero.	11.573	9	.239
Model 2: Grade level SEM			
Preparation			
on PE, KE, & MKT are 0 for all grade levels.	64.843	9	.000
on PE, KE, & MKT are equal for all grade levels.	7.820	6	.252
Practicum length			
on PE, KE, & MKT are 0 for all grade levels.	11.019	9	.274
Practicum quality			
on PE, KE, & MKT are 0 for all grade levels.	13.405	9	.145
Teaching experience length			
on PE, KE, & MKT are 0 for all grade levels.	16.569	9	.056
on PE & KE are 0 for all grade levels.	14.412	6	.025
on PE & KE are equal for all grade levels.	2.262	4	.688
Collaborative activity			
on PE, KE, & MKT are 0 for all grade levels.	9.452	9	.397
Topical experience			
on PE, KE, & MKT are 0 for all grade levels.	18.076	9	.034
on PE, KE, & MKT are equal for all grade levels.	7.337	6	.291
Model 3: Final SEM			
Middle grades and high school indicators equal on PE and equal on KE.	3.508	2	.173
Middle grades and high school indicators equal on MKT.	5.388	1	.020
Preparation, teaching experience length, and topical experience each equal on PE and KE	7.846	3	.049

Table 18

Standardized Path Coefficients and Standard Errors for the Structural Portion of Model 3, a SEM of Mathematical Proficiency for Teaching ($N = 492$)

Predictor	Mathematical proficiency for teaching		
	PE	KE	MKT
	B (SE)	*B (SE)*	*B (SE)*
Certification indicator			
Grades 4 – 8 or EC - 8	0.587 **	.923 ***	0.501 ***
	(0.18)	(0.19)	(0.09)
Grades 8 – 12 or 4 –12	0.746 ***	1.021 ***	0.695 ***
	(0.18)	(0.18)	(0.09)
Perceived academic preparation	0.430 ***	0.670 ***	0.205 ***
for multiplicative reasoning	(0.12)	(0.13)	(0.05)
Recent topical experience with	0.023	0.040	-0.033 **
multiplicative reasoning	(0.02)	(0.02)	(.01)
Teaching experience (years)	0.146 ***	0.107 **	0.036
	(0.04)	(0.04)	(0.03)

** $p < .01$; *** $p < .001$.

Teaching experience was a positive and significant predictor for the factors of TSE, but an increase by 1 year of experience showed an effect size of only approximately 0.10 to 0.15 *SD*. Over 5 years (a period over which novice teachers might come to be called veterans), the effect size became medium to large, ranging from approximately 0.50 to 0.75 *SD*.

Given the evidence discussed above from the Wald tests, the coefficients for grade level as a predictor of TSE were not significantly different between middle and high school certified teachers. The grade level differences in estimates of MKT, however, were significant, and the difference between those estimates ($0.695 - 0.501 = 0.194$) was about the same as the effect size for a one standard deviation increase in perceived academic preparation ($0.194 \approx 0.205$). The Wald tests also showed that the differences between the PE and KE parameter estimates were significant. These parameters show that preparation was relatively more important for KE than

for PE, and that experience was relatively more important for PE than for KE. Finally, the estimate of teaching experience as a predictor of MKT was not significant.

In summary, I found that early entry teachers did not differ from standard teachers on outcomes related to mathematical proficiency for teaching. Contrary to what I had hypothesized, the effects of practicum length and teaching experience on mathematical proficiency for teaching were not significantly different for early entry teachers compared with standard entry teachers. I also hypothesized significant interactions between grade level certification and the effects of preparation, practicum, and experience on mathematical proficiency for teaching. Instead, the relevant Wald tests provided no evidence of significantly worse fit when effects were constrained to equality across certification grade levels. I did find mean differences in the factors of mathematical proficiency for teaching between different grade level certification groups. Rather than the middle grades certified teachers—who are directly responsible for the multiplicative reasoning content—it was the secondary certified teachers who were estimated to have higher means on the measures of mathematical proficiency for teaching (although those differences were not significant for the teaching self-efficacy factors). This finding is concerning but not surprising given research conducted recent at the University of Georgia (Izsák & Jacobson, 2013) and findings elsewhere in the literature on teachers' mathematical knowledge for teaching (e.g., Hill, 2007). The parameters estimated in the final and most parsimonious model of mathematical proficiency for teaching suggest that different outcomes have different antecedents—and thus may develop with some independence.

CHAPTER 4

THE GEORGIA STUDY

This study was designed to address a problem facing administrators and teacher educators working at the district level in Georgia: How can schools and districts in Georgia support teachers' development of mathematical proficiency for teaching? This local policy problem is an instantiation of a national problem that contemporary educational policy seeks to address. Teachers are thought to develop expertise on the job, especially in their first few years of practice, but the conditions that are consequential for learning are not well understood. The ongoing debate of how best to support teachers on the job intersects a related conversation about which teacher characteristics (such as teachers' level of education or their knowledge and beliefs) are related to effective teaching, and thus which characteristics teacher educators and district policies should aim to develop.

Explaining the role of work-related experience in teachers' learning is an important part of research that addresses the problem of school and district support for teacher development, and identifying the relevant aspects of teachers' work-related experiences is an important first step. Several districts in Georgia, inspired by research reporting the positive effects of professional learning communities, have implemented common planning time for grade-level mathematics teachers. Prior research comparing cases of teachers in two different schools suggested that teachers' collaborative work on lesson planning and assessment data can provide opportunities for teachers to develop knowledge useful for teaching mathematics. In one case, a

teacher's collaboration with other grade-level teachers contributed to productive beliefs and goals that supported his professional learning (Jacobson, in press).

The primary goal of this study was to inform local and national policy by exploring the hypothesis that teachers' development of mathematical proficiency for teaching is related to collaboration with other teachers. (As in the Texas study in Chapter 3, I used a mathematical knowledge for teaching, MKT, instrument and a teaching self-efficacy beliefs, TSE Beliefs, instrument to operationalize mathematical proficiency for teaching.) A second goal of the study was to explore why collaboration may support teachers' development in some situations but not in others, with results for practitioners such as administrators and teacher educators. A third goal of the study was to address an open theoretical question concerning the grain size of teachers' conceptual change: Do teachers develop MKT and TSE only for the content topics they teach, or does the mathematical proficiency for teaching developed in one domain of experience transfer to other domains? Under the first hypothesis, one would expect to find change in multiplicative reasoning MKT (for example) only among those teachers who were actively teaching these topics, but if MKT develops at a larger grain size (e.g., teachers learn to apply general principles), then one might see similar change in multiplicative reasoning MKT between teachers who taught different content. Under this second hypothesis, one might also find that multiplicative reasoning MKT was highly correlated with MKT more broadly assessed.

I used measures of mathematical proficiency for teaching in this study that were focused on multiplicative reasoning topics to compare Grade 6 and 7 teachers who teach that content with Grade 8 teachers who do not. The purposes for the study led to the following research questions:

3. What features of teachers' experience in schools are associated with change in mathematical proficiency for teaching multiplicative reasoning topics?
4. How does change in mathematical proficiency for teaching multiplicative reasoning topics differ (a) across schools, (b) between teachers who teach that content and those who do not, and (c) with the frequency of teachers' collaboration and collegial activity focused on student learning?

Data and Methods

The participants in this study were Grade 6, 7, and 8 mathematics teachers in Georgia. They were surveyed three times over the course of the spring semester in 2012. All participants were invited to take the survey on the same date (January 15, March 10, and May 1), and they were allowed to respond to the survey any time in the subsequent 2-week period. Aiming to study change in mathematical proficiency for teaching over a period as short as one semester seemed reasonable to me because teacher educators expect to effect change in prospective teachers' knowledge and beliefs during classes that last only one semester. Because I anticipated that teaching—creating lesson plans, giving explanations, grading assessments, adapting instruction to respond to student learning, and so forth—was educative for teachers under the right conditions (e.g., in collaborative environments), I expected to find measurable change in MKT and TSE over the span of one semester.

I chose to conduct the study during the second semester of the school year to rule out a possible cause of change in MKT and TSE scores that otherwise might have affected the results. It was quite plausible to me that a new teacher (or a newly assigned teacher) of the relevant content—multiplicative reasoning topics—might have substantially higher MKT scores for that content after their first semester as they gained familiarity with the curriculum. If this had

happened then the estimates of teachers' change in MKT would have been biased upwards for new teachers, and I could not have distinguished gains due to teaching experience from those due to familiarization with the curriculum. I wanted to make sure that the baseline measure of teachers' knowledge (Wave 1) reflected teachers' MKT after having some familiarity with the content they were assigned to teach. The curriculum used in Georgia during the 2011–2012 school year included some multiplicative reasoning topics in the fall and spring semesters in both Grades 6 and 7. Multiplicative reasoning topics were not a direct focus of instruction in the Grade 8 curriculum in either the fall or spring semester.

Multilevel modeling methods work best with large data sets, so I made efforts to maximize the amount of survey data collected for this study. Districts in Georgia have local control, so I contacted all districts in Georgia that served 5000 or more students to solicit participation in the study. I did not include very small districts in the solicitation because these districts generally have fewer than 10 Grade 6 to 8 mathematics teachers, and thus there was a low chance that even 1 teacher would participate. In addition, only about 17% of the students in Georgia attended very small districts (see Table 19).

I contacted 74 districts, and 28 agreed to participate, 19 declined, and the rest did not respond. The survey was hosted online. Districts that agreed sent me contact information for teachers, and I emailed these teachers with an identification code to maintain the confidentiality of their responses. In all, I emailed 799 teachers in the 14 small and 10 medium-sized districts that agreed to participate. I additionally emailed 198 teachers in the two large districts. Of these 997 teachers, 329 teachers began taking the first survey. The 199 teachers who completed the first survey formed the analytic sample for this study (a 20% response rate). The teachers in the analytic sample came from 85 schools and 26 districts. Although this sample was not a

representative probability sample, it is likely that the variability in the analytic sample captured much of the variability in the state because the sample included teachers from a large number of different schools and districts.

Table 19

District Participation in the Study

	Districts in Georgia				Of agreeing districts		
	Total	% of students served	No reply or not eligible	Denied	Agreed	Participated	% of students served
Large districts (>90k students)	5	27.3	1	0	4	2	15
Medium districts (11k to 51k students)	28	40.7	5	13	10	10	13
Small districts (11k to 5k students)	34	15.2	14	6	14	14	6
Very small districts (<5k students)	125	16.7	125	-	-	-	-
Total	192	100	139	19	28	26	34

Table 20 describes how response rates were distributed by survey wave. The teachers in the analytic sample responded to three surveys available online on January 15 (Wave 1), on March 9 (Wave 2), and on May 1 (Wave 3). In general, respondents replied within 2 or 3 weeks and none took longer than a month; the last survey was received on May 15. In an effort to reduce attrition, I used 1 email forewarning, 1 email request, and 3 email reminders for each of these waves. There were 149 teachers who completed the second survey, and 132 who completed the third survey. Only 35 teachers in the analytic sample took only one survey (the first), and 47 others took only two surveys (the first and either the second or the third).

Table 20

Number of Participants in the Analytic Sample Who Did and Did Not Participate in Different Survey Waves

	Wave			
	1	2	3	2 & 3
Participated	199	149	132	117
Did not participate	0	15	32	35

I was primarily interested in discovering how districts might better support beginning teachers, so I selected a subset of participants based on years of teaching experience to interview. Of the 199 survey participants, I found 35 who had 5 or fewer years of experience. I contacted all 35 and asked them to participate in interviews. Of the 35, 8 did not respond, 6 declined to participate, and 21 agreed to participate. Two of the 21 who agreed to participate did not respond to emails for scheduling the interview, and another one opted to answer the survey questions by email rather than face-to-face but did not reply to follow up emails. I conducted interviews with the remaining 18 teachers. The data from one interview—with a first-year special education teacher from a small district—were unusable because of a technical problem with the recording. The data reported for this study include 480 surveys from 199 participants and 17 audio-recorded interviews, each about 1 hour in length.

Instruments

All three surveys included an instrument for measuring mathematical knowledge for teaching (MKT) and a teaching self-efficacy (TSE Beliefs) instrument. The MKT instrument used for this study was the short MKT instrument (17 items) described in Appendix B. The TSE Beliefs instrument used for this study was very similar to the instrument used in the Texas study and described in Chapter 3 and in Appendix B. The difference is that I changed each item from

asking about "topics involving fractions, ratios, and proportions" to asking instead about "topics involving ratios and proportions." I made this change to ensure a better fit between the TSE Beliefs measure and the experiences of the teachers participating in the study. None of the curriculum topics in 6th, 7th, or 8th grade in Georgia in the year of this study focused explicitly on fraction topics, but all three grade levels made use of ratios or proportions to some extent. Direct instruction of ratio and proportion topics was limited to the 6th and 7th grades. I compared the fit of a one-factor and two-factor confirmatory factor analysis models with data from the revised TSE Beliefs instrument and found strong evidence that the revised TSE Beliefs instrument was working as expected. In particular, the TSE Beliefs instrument was composed of two factors: personal efficacy (PE) and knowledge efficacy (KE).

The first survey included several questions that addressed the participants' background, preparation, and school contexts, including questions about the grade level they currently. The third survey asked about professional learning activities (such as coaching and hours of mathematics-related professional development) and the teachers' experiences over the spring semester including their (1) frequency of activity with colleagues focused on student assessment data, instruction, and lesson planning (see Chapter 2 for more information about this measure), (2) whether or not they had experience teaching (or extensive tutoring time) with multiplicative reasoning topics such ratio and proportion topics.

The variables used in the analysis included categorical indicators for gender, ethnicity, early entry certification route, and Grades 6–12 credential. I also used a categorical indicator for whether teachers had relevant topical experience in the first semester (i.e., they taught ratio and proportion topics). In addition, I used the following continuous variables: teaching experience measured in years, teachers' self-reports of their number of hours of mathematics related

professional development for each semester, and a scale score summarizing 5 Likert items concerning collaborative activity for each semester. Summary statistics for these variables are provided in Table 21.

Table 21.

Summary Statistics for the Explanatory Variables.

	Percentage
Gender (male)	18.1
Ethnicity	
Asian	0.5
Black	13.1
Hispanic	1.0
White	83.4
Other or no response	2.0
Certification route (early entry)	26.6
Credentialed for Grades 6–12	23.1
Grade level [a]	
Grade 6	37.2
Grade 7	38.7
Grade 8	41.2
Topical experience	76.4
	Mean
Years of experience (years)	10.6
Math PD – semester 1 (hours)	8.1
Math PD – semester 2 (hours)	9.9
Collaboration [b] – semester 1	4.0
Collaboration [b] – semester 2	3.8

[a] Grade level percentages do not add to 100 because many teachers were assigned to more than one grade.
[b] The scale for collaboration ranges from 1 (*never*) to 5 (*once each week*).

The continuous variables were recoded (or summed, in the case of the collaboration variable) before analysis. Teaching experience was reported in years, and recoded for analysis on an 8-point scale: <1, 1–2, 3–5, 6–10,11–16, 17–24, 25–32, >32 years. The topical experience variable came from a list of topics teachers were asked to check if they taught those topics or spent an extensive amount of time tutoring students in those topics. The variable used in analysis

indicated whether or not teachers checked the box for "ratio and proportion." Mathematics professional development was measured in total hours and recoded on an 8-point scale: <1, 1–2, 3–5, 6–10, 11–16, 17–24, 25–32, and >32 hours.

The collaboration variable was included in the analysis to reflect differences in teachers' professional work environment. This variable was modeled in analyses as a latent factor with five indicators (α = .78), each a 5-point scale (1: never; 2: once a year; 3: once or twice each semester; 4: once a month; 5: once each week). These five indicators asked about the frequency of professional activities that might support teachers' learning with colleagues, especially activities focused on student thinking. Teachers were asked how frequently in the past 3 years they had done each of the following activities with colleagues: analyzed sample student work, sought advice about instructional issues, discussed teaching practice, discussed the strengths or needs of specific students, and discussed student assessment data to make instructional decisions.

Retrospective interview protocol. The 18 interviews conducted for this study followed a semi-structured interview protocol (Kvale, 2007; Seidman, 2006; see Appendix C). The participants were asked to describe various aspects of their teaching practice, such as the use of manipulatives, and how these aspects of practice had changed from their first year of teaching. To assess teaching self-efficacy, I asked the participants about their confidence for teaching mathematics (personal efficacy, PE) and their confidence in their knowledge of the mathematics they used in teaching (knowledge efficacy, KE). I also asked how their confidence had changed since they began teaching.

The format of the interview balanced an open-ended aspect for eliciting descriptions true to the experiences of the participant, on the one hand, with a deliberate focus on how change in mathematical proficiency for teaching might be experienced, on the other. The semi-structured

nature of the interview protocol gave me opportunities to follow up on surprising statements that I perhaps could not have anticipated before the interview and helped yield data that complemented the survey data. For example, one of the first teachers I interviewed contrasted her first-year experience of a mentor teacher who taught in a different grade with her experience in her second year of a mentor teacher who taught in the same grade. She found the same-grade pairing critical for the mentoring to be helpful. From that point on, I asked all of the teachers who described mentor teachers about grade level similarities and differences.

Qualitative Analysis

In this study, I emphasized quantitative analysis, and I used qualitative data to augment and help interpret the quantitative findings. I transcribed 8 of the 17 interviews verbatim and had a transcription service produce initial transcripts for the remaining 9 interviews. I listened to and revised these externally produced transcripts to verify and improve their accuracy. Then I organized transcript data from participant interviews into two documents to find patterns across interviewees about (1) how teaching practice had changed from the first year until the interview, and (2) how MKT, PE, and KE had changed over the same time period. The patterns I noticed across participants helped me better understand the context of the results I obtained from analyzing the survey data. At the end of each interview, I asked about the most important factor that had contributed to their professional growth, and teachers' responses to this question helped answer research question 3. I also had asked interview questions about each teacher's collaborative work with colleagues or mentors, how that work had changed over time, and the impact it had had on their professional growth. These data helped me answer research question 4.

I had hoped that I could steer the conversation during these interviews to focus on multiplicative reasoning topics so that I would be able to better understand the change in mathematical proficiency for teaching multiplicative reasoning topics. In some interviews, I was able to do this, but in most I was not. Overall, the interview data reflected the teachers' accounts of their changing mathematical proficiency for teaching, but at a coarser grain size than the domain of multiplicative reasoning. The descriptions of changing practice provided evidence of changes in teachers' mathematical knowledge for teaching, particularly their knowledge of curriculum, tasks, and representations. I also asked about how the teachers' PE and KE changed, and looked across the set of responses to find common themes.

Quantitative Analysis

I used the *lmer* method in the R package *lme4* to analyze the survey data, and used a general multi-level model with three levels. The first level describes individual change over time, the second explains how individual change differs between participants, and the third level describes how participants differ between schools (Hox, 2010; Singer & Willett, 2003).

The first level of the model I used in this study hypothesized linear growth in participants' MKT, PE, and KE over time. It is quite possible that growth in MKT is not actually linear. Research with children in the Piagetian tradition (e.g., Steffe & Olive, 2010) has shown that proficiency with problem solving in the domain of multiplicative reasoning can be rapid as students apply similar ways of thinking across a wide range of new problems with similar mathematical structures through assimilation, and that there are long plateaus wherein students struggle to accommodate existing mental operations to solve new problems with novel mathematical structures. On the other hand, research on expertise suggests that time and experience under the right conditions are related to the development of expertise (Ericsson,

2004). Change in beliefs about mathematics and teaching can be similarly complex (Philipp, 2007). The linear model used in this study is a reasonable first approximation, and more complex, curvilinear models of change should be explored in future studies. (An investigation of nonlinear models of change requires more than three survey waves and thus was not possible with the data collected for this study.)

Under the assumption of linearity at Level 1, different versions of the model can be obtained by including different variables at the second and third level, such as a participant-level (Level 2) variable reflecting the frequency of collegial activity or a school-level (Level 3) variable reflecting average collegial activity to explain differences in individual change over time. Two models are called *nested* if one has all the variables in the other plus additional variables. Hypothesis tests exist for comparing nested models, and those tests allowed me to determine whether models that included key variables (whether or not teachers taught ratio and proportion; whether or not teachers worked with mentors or colleagues) fit the data better than simpler models without those variables. The final models for MKT, PE, and KE provided estimates of the effects and standard errors of key variables on the outcome variables. Selection bias remained a limitation in this analysis in that teachers who varied on the variables of interest (e.g., mentoring or teaching ratio and proportion) might also have systematically differed on characteristics that affect MKT, PE, and KE. Known predictors of MKT (e.g., Hill, 2007, 2010) were included in the models as covariates to mitigate possible bias, but unobserved covariates might have remained. The results must be interpreted as descriptive rather than causal.

Missing Data

One current best practice for handling missing data is to use the method of multiple imputation (Enders, 2010). In this method, many data sets are created that contain all observed

data and plausible values (imputations) for the missing data; then each data set is analyzed, and the results are combined (Rubin, 1987). Because the imputed values differ across the different data sets, the variability that observed data would have had is reconstructed for the analysis. Imputation works well when there are many observed variables for an individual and only a few missing values—the observed data are used to predict (with appropriate variability) values to replace the missing data. Longitudinal data created further complications because some variables are time varying (the outcomes vary within individuals over time), but others are time invariant. In the present study, data on time-invariant variables were collected on the first or third surveys, and because of attrition, time-invariant variables collected on the third survey had a greater percentage of missing values than those collected on the first survey.

Table 22

Percentage of Missing Data on the Explanatory Variables

Explanatory variable	Wave 1 participants (n = 199)	Wave 3 participants (n =132)
Gender	2.0	-
Race/ethnicity	1.5	-
Early entry	1.5	-
6–12 credential	1.5	-
Grade level	0.5	-
Years of experience	1.0	-
Topical experience	1.0	-
Math PD – Semester 1	1.5	-
Collaboration – Semester 1	0.5	-
Math PD – Semester 2	39.2	8.3
Collaboration – Semester 2	41.7	12.1

I used a three-step process recommended by Gelman and Hill (2007, p. 541) to impute the missing data for this study. The goal was to create 50 complete versions of the data set, each representing a plausible data set that might have been collected if all 199 participants had taken all 3 surveys (597 surveys in all) and answered all of the questions on each survey. In fact, only

480 surveys were completed, and a substantial portion of the data from individual surveys were missing because some participants did not complete the whole survey. The MKT, PE, and KE outcome variables were missing data on 6%, 2%, and 3% of the completed surveys (n = 480), respectively, and they were missing data on 24%, 21%, and 22% of the 597 administered surveys, respectively. Table 22 reports the percentage of missing data on the predictors; the questions about second semester professional development and collaboration activity were on the third survey.

To make the final 50 imputed data sets used for analysis, I first created 5 data sets with imputed values for each participant's outcome variables on each survey (597 values) using the R software package Amelia II. This package uses a regression model and expectation-maximization algorithm to predict missing values using observed data. The initial imputation step used school climate (not otherwise used in this analysis) and collaboration items for which there was complete response data. I used each of the 5 initial data sets to calculate individual MKT, PE, and KE means for the 199 participants. Then I used each set of outcome means and the observed data on the predictors to create 5 new data sets with imputed values for each of the 199 participants. Each of the resulting 25 imputed data sets had complete imputed data on the predictors, and I used each to create 2 new imputed data sets with reimputed MKT, PE, and KE values. The resulting 50 data sets were used for analysis.

Results

In this section, I address each research question in turn with quantitative methods at both the individual and school level. I conclude by drawing on findings from the interviews to provide a context for and to help making sense of the results of the quantitative analysis.

Quantitative Results

I began the analysis of the quantitative survey data by using multilevel models with no predictors to examine the variation in the three outcome variables: mathematical knowledge for teaching (MKT) and the two factors of teaching self-efficacy beliefs, personal teaching self-efficacy (PE) and knowledge self-efficacy (KE). The goal of this analysis was to determine how the observed variation in scores was partitioned within and between the individual (Levels 1 and 2) and between schools (Level 3). That is, I found how much overall differences between teachers initial MKT, PE, and KE could be attributed to the fact that they were working in different schools.

Table 23

Median Variance and Variance Partition Coefficients for the Individual and School Levels in 2- and 3-Level Models of MKT, PE, and KE Across 50 Imputed Data Sets

Outcome	Model	Residual variance	Individual variance	School variance	Individual VPC	School VPC
MKT	2 levels	.207	.351	-	.629	-
	3 levels	.207	.338	.0108	.608	.019
PE	2 levels	.406	.574	-	.586	-
	3 levels	.406	.551	.0272	.560	.028
KE	2 levels	.341	.592	-	.634	-
	3 levels	.341	.538	.0536	.577	.057

Note. Residual variance is all the variance not accounted for by the higher levels and in particular includes within individual variance over time.

I found that little variation in any of the outcomes was accounted for at the school level. The variance partition coefficient (VPC) describes the portion of overall variance observed at the specified level of the model. A school-level VPC of less than .02 for the MKT outcome meant that less than 2% of the variance in the teachers' MKT scores was accounted for by random school-level effects (see Table 23). The percentage of variance in PE and KE at the school level

was higher but still relatively small compared with the overall variance in those outcomes; less than 3% of the variance in PE and less than 6% of the variance in KE was accounted by school-level random effects. These findings mean that across the data set, the outcomes of the teachers in the same school were not much more similar than the outcomes of teachers in different schools. In contrast, I found that random effects at the individual level accounted for a large percentage of the variance in each outcome. The individual VPCs were greater than 50% for all of the outcomes: The individual VPC for MKT was 61%, for PE it was 56%, and for KE it was 58%. These results confirm that (on average and across the data set) two measurements of an outcome from the same individual were much more similar than two measurements of that outcome from different individuals.

The results about how variance in the outcomes was partitioned between individuals and schools justified the use of 2-level longitudinal models for these data. These results explained why 2-level models fit these data much better than either single-level regression models (which constrain all individuals to have the same initial value and rate of change) or 3-level models, which estimate common intercepts and slopes for teachers in the same school. In the rest of the analysis, I focus on results from 2-level models of outcomes that estimate separate initial values and rates of change for each individual. For each model reported below that includes predictors, I also fit analogous single and 3-level models (which are not reported), and in each case, the fit of the 2-level model was significantly better than either of the others.

The analysis presented thus far supports the conclusion that that the features of teachers' experience in schools have little differential effect by schools (3rd research question) and that initial values of outcome and rates of change in mathematical proficiency for teaching differs little across schools (4th research question). To address these research questions (both the

relationship between teachers' school experiences and their mathematical proficiency for teaching and how change differs between teachers) when individual teachers are the unit of analysis, I next fit a sequence of four models for each outcome variable. Each sequence included two baseline models recommended by Singer and Willet (2003) that did not have any predictors.

The first baseline model is called the unconditional means model (Model 1) and predicts the outcome as the sum of an intercept term (the grand mean of the outcome) and a random error term, both at the measurement occasion level (Level 1) of the model. The variance of the random error term in Model 1 captures the variation of the outcome across individuals and measurement occasions. The unconditional growth model (Model 2) predicts the outcome as the sum of an intercept term and a random error term at the measurement occasion level (Level 1) and an intercept term and a random error term for time at the individual level (Level 2). The variance of the random error term for time captures the variation in growth rates across individuals in the study.

The unconditional means model (Model 1) and the unconditional growth mode (Model 2), provide useful comparisons with the models that include predictors (Models 3 and 4). Moreover, by comparing Model 1 and Model 2 and calculating the percentage decrease in residual (Level 1) variance, one can determine the amount of variation explained by time. Time accounted for .062 of the within-individual variance in the MKT outcome, .074 of the within-individual variance in the PE outcome, and .261 of the within-individual variance in the KE outcome.

The unconditional growth model also provided baseline estimates of the rate of change for each outcome over the period of the study. The estimated rate of change was positive in the model of MKT and estimated to be 0.03 *SD* per month or about 0.12 to 0.13 *SD* over the

semester. The estimated rate of change for the teaching self-efficacy beliefs outcomes were both negative, and estimated to be -0.06 and -0.07 *SD* per month for PE and KE, respectively, or about -0.22 and -0.26 *SD*, respectively, over the semester. In Model 2 for all three outcomes, the intercept (outcome at time zero) and the rate of change in the outcome associated with time were moderately and negatively correlated (-.43 or -.44), which means that higher outcomes at the beginning of the study were associated with slower growth in MKT and a faster decrease in PE and KE. These findings are noteworthy because they indicate that measureable change in knowledge and beliefs did occur within the 4-month period of the study and that the effect size for one semester of teaching experience was small to moderate, between 0.10 and 0.30 *SD* on the outcomes used in this study.

Next, I consider each outcome in turn and address the research questions by reporting how and to what extent experience teaching multiplicative reasoning topics and collaborative activity focused on student learning were related to the outcomes. I also consider the teachers' background and experience in schools more generally and report on how features of background and experience were related to initial outcomes at the beginning of the study and to change in the outcome measures over the course of the study.

Predictors of initial status and change in MKT. Table 24 reports the parameter estimates for the four models of MKT. Model 4 includes all predictors examined in the study, and Model 3 is a more parsimonious model that includes just those predictors that were statistically significant. Model 1 and Model 2 are the unconditional means and unconditional growth model and have already been introduced.

Several statistics are useful for understanding how well these models fit the data. In regression analyses, the R^2 statistic can be interpreted as the fraction of variance explained and

can be computed by squaring the predicted and observed outcome variable. In multilevel models, there is not a single statistic with the same interpretation, and many so-called pseudo R^2 statistics are used. The $R^2_{y\hat{y}}$ statistic reports the squared correlation between the model-predicted and observed outcomes. This value was already high for the unconditional means model (.75) and increased slightly for Models 2, 3, and 4 (.77 to .78). The increased number of predictors in the last two models did not change this measure of fit very much. Two other R^2 statistics are reported in Table 24, and these have a different meaning: they describe the fraction of variance in the initial status (at time equals 0) and slope (rate of change associated with time) in the unconditional growth model that is explained by the added predictors. A little more than 11% and 12% of the variance in initial status MKT was explained by Model 3 and Model 4, respectively. In addition, 50% and 41% of the slope for MKT was explained by the predictors in Model 3 and Model 4, respectively. One of the curious features of multilevel models is that, unlike in regression models, the amount of variance explained does not always go down when more predictors are added (Singer & Willet, 2003, p. 104). The reason for this surprising behavior is that predictors added at one level can affect the variance in other levels of the model. Such behavior is especially common in models—like the ones reported in this study—where most of the variation is between individuals rather than more evenly balanced within and between individuals.

Table 24

Models of Mathematical Knowledge for Teaching (MKT)

	Model			
Variable	1	2	3	4
Fixed effects				
Initial status				
Intercept	0.055 (.048)	-0.009 (.056)	-0.236 (0.164)	-0.234 (0.177)
Male	-	-	-	0.126 (0.124)
Nonwhite	-	-	-0.463 *** (0.130)	-0.495 *** (0.148)
6–12 credential	-	-	0.254 * (0.107)	0.237 * (0.109)
Experience	-	-	0.071 * (0.031)	0.072 * (0.031)
Math PD	-	-	-	-0.011 (0.020)
Collabor. (COL)	-	-	-0.192 * (0.098)	-0.183 (0.101)
Topics (TOP)	-	-	-0.073 (0.110)	-0.058 (0.114)
COL × TOP	-	-	0.179 [b] (0.109)	0.182 (0.110)
Slope				
Intercept	-	0.031 (.016)	0.051 * (0.023)	0.057 (0.032)
Nonwhite	-	-	-	0.017 (0.043)
Early entry	-	-	0.066 * (0.030)	0.062 * (0.031)
Math PD	-	-	-0.011 * (0.006)	-0.011 (0.006)
COL	-	-	-	0.005 (0.025)
Grade 6, 7 (GR67)	-	-	-	-0.012 (0.029)
COL × GR67	-	-	-	-0.002 (0.028)
Random effects [a]				
Level 2				
Var. - Initial status	0.351	0.396	0.351	0.348
Var. - Slope	-	0.0032	0.0016	0.0019
Corr. - Initial-slope	-	-.43	- .56	- .52
Level 1				
Residual variance	0.207	0.194	0.193	0.193
Goodness-of-fit [a]				
Initial status R^2	-	-	.11	.12
Slope R^2	-	-	.50	.41
$R^2_{y\hat{y}}$	.75	.78	.77	.77
Log-likelihood	-558	-559	-554	-569
AIC	1122	1129	1137	1178
BIC	1135	1155	1186	1265

Note. Standard errors for the regression coefficients are given in parentheses.
* p < .05; ** p < .01; *** p < .001.
[a] The random effects and goodness of fit estimates reported are the median estimate across the 50 imputed data sets used for this study.
[b] The *p*-value for this estimate is .102.

I next discuss the relative fit of the models to provide context for interpreting the estimated model parameters. To assess relative model fit, I computed log-likelihood ratio tests between each model fit to each of the 50 imputed data sets. There are standard rules (e.g., Rubin, 1987) for combining parameter estimates and standard errors across models, but none for combining the log-likelihood test statistics. There is evidence from a simulation study (Asparouhov & Muthén, 2010) that the mean of these test statistics is biased. I therefore report the median as well as the 1st and 3rd quartiles to give a sense of the range of test results across the set of 50 imputed data sets (see Table 25). This range is not a confidence interval, but it does reflect the variability between the imputed data sets and can support the assessment of model fit.

The tests between Model 1 and Model 2 for the MKT outcome reveal mixed support for including time as a random effect. The likelihood ratio test (LRT) statistic has three degrees of freedom because Model 2 specifies two new variance parameters for the random effects associated with the intercept and slope of time and a parameter for the covariance. Variance parameters cannot be negative, and the null hypothesis for this test constrains the two variance parameters to zero. This is the boundary of the parameter domain, and thus the test statistic is a mixture of different chi-square distributions and does not follow $\chi^2_{df=3}$. The reported *p*-value in Table 25 for the Model 1 vs. Model 2 test is an upper bound on the possible *p*-value of the test. Thus, the finding that half of the models have upper bounds on *p*-values that are close to statistical significance at the .05 level suggests that this test might yield a significant value. Certainly the null hypothesis (that setting the variance and covariance parameters for the intercept and slope of time to zero does not reduce model fit) is not tenable in at least a quarter of the imputed data sets.

I next considered two other criteria of model fit. Statistically preferred models in a set have a smaller Bayesian information criterion (BIC) and a smaller Akaike's information criterion (AIC). These criteria can used to compare the fit of nonnested models, but are also useful for assessing the fit of nested models. Table 25 shows that Model 2 has a slightly lower increase in AIC but a higher increase in BIC. These indexes both penalize extra parameters to favor parsimony, but the way parameters are penalized differs and can sometime result in conflicting signals about fit. On balance, the mixed information on model fit led me to retain the random effects for the intercept and slope of time.

Table 25

Comparisons of MKT Models Over 50 Imputed Data Sets

Comparison	LRT	*df*	*p*	Increase in AIC	Increase in BIC
Model 1 vs. 2 [a]					
3rd quartile	9.38	3	.025	-3	9
Median	7.12	3	.068	-1	12
1st quartile	5.02	3	.171	1	14
Model 2 vs. 3					
3rd quartile	48.8	8	< .001	-33	2
Median	42.6	8	< .001	-27	8
1st quartile	37.3	8	< .001	-21	14
Model 3 vs. 4					
3rd quartile	4.35	6	.629	8	34
Median	2.55	6	.863	9	35
1st quartile	1.90	6	.928	10	36

[a] Testing variance components involve the domain boundary (variance cannot be less than zero), and so the distribution of the test statistic does not follow $\chi^2_{df=3}$; the reported *p*-value is an upper bound for the true *p*-value of the test.

Testing Model 2 against Model 3 (the model that included all significant predictors) had clearer results. Although the BIC increased somewhat across the majority of the imputed data sets, there was strong evidence from the *p*-values of the LRT for rejecting the null hypothesis (no loss of fit with Model 2) and retaining Model 3: the tests across all 50 data sets were highly

significant. (Model 3 also had a statistically significantly better fit than Model 1.) Finally, I tested Model 3 against Model 4, but found no reason to prefer the more complex model.

I next interpret the results of the parameters for Model 3. The statistically significant and large coefficient for the nonwhite indicator was surprising. No other research on MKT of which I am aware has reported significant differences in MKT related to participants race or ethnicity. The term represents a comparison between 165 self-identified white teachers in the sample and a heterogeneous group of 34 others (26 black, 2 Hispanic, 1 Asian, and 4 others). Analysis of further subgroups was not possible because of the size of the subgroups severely limited power. This result has several possible meanings, but none make much sense to me. The result could reflect true differences in knowledge, but these differences have not been observed in other research and I did not find them in the study of Texas teachers using a very similar MKT measure. The result could also reflect selection bias into the study. It is possible, for example, that only high knowledge white teachers participated in the study. However, the question about race was the second to last question on the survey, so any effect like this should have been evident in different rates of attrition as well, but this was not the case. The result could also reflect selection bias into the teaching profession, but only if we accept the possibility that race and teacher hiring in Georgia are related by a relationship that is not evident in other states, including Texas. A third explanation is provided by the hypothesis of differential item functioning (DIF); something about the wording of some questions may have made them easier for white participants. This possibility also lacks plausibility because most of the items (all but the 2 items from the DTMR project) were screened for DIF by the Educational Testing Service, and all the items were administered in the Texas study and no DIF related to the race variable was observed. At the same time, there was not enough information in the data set to rule out any

of these hypotheses. (The limited number of participants in the Georgia study precludes DIF analysis.)

As I expected from prior research identifying predictors of MKT (e.g., Hill, 2007, 2011), the indicator for a Grades 6–12 credential was a significant predictor of MKT at the beginning of the study. Teaching experience was also a significant predictor of initial MKT. Collaborative activity was measured at the beginning and end of the survey, and participants were asked to report their collaborative activity around student learning for the first and second semesters, respectively. The topics variable indicated which teachers reported teaching ratio and proportion topics in the school year. The significant negative parameter estimate for collaborative activity indicates that teachers who regularly worked with others to plan lessons, evaluate student work, and the like had lower MKT at the beginning of the study than teachers who did not engage in such activities. Similarly (although the parameter was not statistically significant), the model described teachers who taught ratio and proportion topics as having lower MKT than those who did not. What is most interesting about these two parameters, however, comes in their interaction. Collaborative activity had a positive (but not significant) effect on MKT for those teaching ratio and proportion topics. If the interaction term is removed from the model, then neither collaborative activity nor ratio and proportion topics experience is a significant predictor of initial MKT.

I found that collaborative activity had no significant effect on the slope of MKT, contrary to my hypothesis. Moreover, the Grades 6 and 7 teachers—who were responsible for teaching ratio and proportion and other multiplicative reasoning topics—did not increase their MKT at different rates than Grade 8 teachers. There was no significant interaction between grade level and collaborative activity on the estimated rate of change for MKT. Instead, I found that the

early entry teachers (those who had begun teaching before completing full licensure) learned MKT significantly faster over the semester than the teachers who were traditionally certified. Mathematics professional development during the second semester was negatively associated with the rate of change in MKT, with slower growth for those who were involved in more mathematics professional development. In the sample, 41 teachers did not participate in mathematics professional development during the second semester, and of those who did, the median score was 4 or about 11 to 16 hours.

The estimated rate of change for MKT in Model 3 was higher than in Model 2 at 0.05 standard deviations per month, or 0.2 standard deviations over the semester. Among the 53 early entry teachers, the average rate of change more than doubled, to 0.12 standard deviations per month or approximately 0.5 standard deviations over the semester. Most of these teachers (42), however, were involved in professional development during the second semester, and thus the predicted growth rate was lower. In all, for 9 early entry teachers, the positive effects of early entry and the negative effects of mathematics professional development cancelled out (a score of 6 on the mathematics PD scale). The 2 early entry teachers who participated in more than 24 hours of professional development (scores of 7 and 8 on the mathematics PD scale) saw an overall negative effect on their MKT rate of change from these two variables.

Predictors of initial status and change in personal efficacy (PE). Table 26 reports the comparisons of fit between the four models of PE, and Table 27 reports the goodness-of-fit statistics and parameter estimates for these models. The test directly comparing Model 1 and Model 2 (see Table 26) was significant for more than 75 of the imputed data sets (the upper bound for the *p*-value was less than .05). As with the analogous models for MKT, the AIC and BIC criteria gave conflicting information about which model was statistically preferred, with

decreases in the AIC but increases in the BIC associated with Model 2. The test between Model 2 and Model 3 provided strong evidence that Model 3 was preferred (although BIC increased slightly). The test between Model 3 and Model 4 showed that the more parsimonious model did not have worse fit than the model with more predictors. The $R^2_{y\hat{y}}$ statistic increased from .72 to .76 in the three models that included time and other predictors (see Table 27), indicating that observed and predicted values of PE were more highly correlated in Models 2, 3, and 4. Models 3 and 4 both explained a substantial portion of the variance in initial PE status and the slope for PE—more than 20. Together, these results indicate that the models of PE that include time and other predictors more accurately reflect the data than the unconditional means models.

Table 26

Comparisons of Personal Efficacy (PE) Models Over 50 Imputed Data Sets

	LRT	*df*	*p*	Increase in AIC	Increase in BIC
Model 1 vs. 2 [a]					
3rd quartile	12.9	3	.005	-7	6
Median	9.7	3	.021	-4	9
1st quartile	7.9	3	.048	-2	11
Model 2 vs. 3					
3rd quartile	62.2	10	< .001	-42	2
Median	58.0	10	< .001	-38	6
1st quartile	54.5	10	< .001	-34	9
Model 3 vs. 4					
3rd quartile	4.4	4	.238	2	20
Median	2.6	4	.518	5	22
1st quartile	1.9	4	.757	6	24

[a] Testing variance components involves the domain boundary (variance cannot be less than zero), and so the distribution of the test statistic does not follow $\chi^2_{df=3}$; the reported *p*-value is an upper bound for the true *p*-value of the test.

Table 27

Models of Personal Efficacy (PE)

Variable	Model 1	2	3	4
Fixed effects				
Initial status				
Intercept	0.003 (.060)	0.115 (0.076)	-0.911*** (0.213)	-0.940*** (0.222)
Male	-	-	-	0.118 (0.158)
Nonwhite	-	-	-0.325 * (0.161)	-0.542 ** (0.191)
6–12 credential	-	-	-	0.128 (0.137)
Experience	-	-	0.188*** (0.039)	0.194*** (0.040)
Math PD	-	-	0.064 ** (0.024)	0.062 * (0.025)
Collabor. (COL)	-	-	-0.161 (0.126)	-0.142 (0.128)
Topics (TOP)	-	-	0.118 (0.142)	0.108 (0.143)
COL × TOP	-	-	0.281 * (0.136)	0.275 * (0.137)
Slope				
Intercept	-	-0.055 * (0.022)	-0.067 * (0.034)	-0.092 * (0.044)
Nonwhite	-	-	-	0.105 (0.058)
Early entry	-	-	-	0.005 (0.041)
Math PD	-	-	-	-0.001 (0.008)
COL	-	-	0.085 * (0.034)	0.084 * (0.036)
Grade 6, 7 (GR67)	-	-	0.015 (0.038)	0.029 (0.039)
COL × GR67	-	-	-0.100 ** (0.038)	-0.099 * (0.039)
Random effects [a]				
Level 2				
Var. - Initial status	0.574	0.667	0.514	0.516
Var. - Slope	-	0.0072	0.0051	0.0054
Corr. - Initial-slope	-	- .44	- .45	- .44
Level 1				
Residual variance	0.406	0.377	0.374	0.374
Goodness-of-Fit [a]				
Initial status R^2	-	-	.23	.23
Slope R^2	-	-	.25	.29
$R^2_{y\hat{y}}$	.723	.762	.755	.754
Log-likelihood	-744	-741	-733	-739
AIC	1494	1495	1498	1519
BIC	1507	1521	1568	1606

Note. Standard errors for the regression coefficients are given in parentheses.
* p < .05; ** p < .01; *** p < .001.
[a] The random effects and goodness of fit estimates are the median estimates across the 50 imputed data sets.

In Model 3 for PE (as in Model 3 for MKT), the coefficient for the nonwhite indicator was statistically significant and positive, and that finding was surprising for the same reasons as those discussed for Model 3 of MKT. Additionally, teaching experience was a statistically significant predictor of the initial PE, along with mathematics professional development. Both of those predictors were positively associated with initial PE. The interaction between collaborative activity and teaching experience with ratio and proportion topics was statistically significant, and was positively related to initial PE. Collaborative activity was negatively related to initial PE (but that relationship was not significant), and the indicator for teaching ratio and proportion topics was positively related to initial PE (also without statistical significance). These relationships may have substantive significance even though they are somewhat ambiguous because the *p*-values for each were above .05. Model 3 suggested that the teachers who collaborated would have lower initial PE for multiplicative topics, but if the collaborating teachers also taught ratio and proportion topics, then they were predicted to have higher initial PE than the initial PE of those who taught ratio and proportion but did not collaborate.

Next, I discuss the Model 3 predictors of the rate of change in PE. The overall estimate for the rate of change was negative, a predicted loss of about 0.07 *SD* per month. The estimate of the effect of collaborative activity on the rate of change was significant and positive. Grades 6 or 7 teaching experience was not significantly related to the rate of change in PE. The interaction term between collaborative activity and the indicator for teaching Grade 6 or 7 was significant, but it was negative. Together, these estimates mean that the teachers who taught Grade 8 only and were more involved in collaborative activity had a much slower decrease in PE than the other teachers (and actually a predicted increase of almost 0.02 *SD* per month for those 1 *SD* above the mean on the collaborative activity scale), but that the Grades 6 and 7 teachers who

worked with others frequently had a much faster decrease (almost twice as fast) as the Grades 6 and 7 teachers who worked less frequently with other teachers.

Predictors of initial status and change in knowledge efficacy (KE). Table 28 reports the comparisons of fit between the four models of KE, and Table 29 reports the goodness-of-fit statistics and parameter estimates for these models. The tests of model fit reported in Table 28 provide strong evidence that Model 2 significantly outperformed Model 1 and that Model 3 significantly outperformed Model 2 in terms of model fit. Both of these comparisons show decreases in AIC and BIC in addition to highly significant upper bounds for *p*-values of the tests across the majority of the imputed data sets. Model 4 did not fit the data better than Model 3.

Table 28

Comparisons of Knowledge Efficacy (KE) Models Over Imputed Data Sets

	LRT	df	p	Increase in AIC	Increase in BIC
Model 1 vs. 2 [a]					
3rd quartile	30.9	3	< .001	-25	-11
Median	27.1	3	< .001	-21	-8
1st quartile	21.3	3	< .001	-15	-2
Model 2 vs. 3					
3rd quartile	36.9	4	< .001	-29	-11
Median	34.8	4	< .001	-27	-9
1st quartile	32.8	4	< .001	-25	-7
Model 3 vs. 4					
3rd quartile	12.7	10	.242	7	51
Median	10.8	10	.374	9	53
1st quartile	9.5	10	.485	10	54

[a] Testing variance components involves the domain boundary (variance cannot be less than zero), and so the distribution of the test statistic does not follow $\chi^2_{df=3}$; the reported *p*-value is an upper bound for the true *p*-value of the test.

Table 29

Models of Knowledge Efficacy (KE)

Variable	Model 1	2	3	4
Fixed effects				
Initial status				
Intercept	0.006 (.060)	0.142 (.076)	-0.751*** (0.187)	-0.922*** (0.227)
Male	-	-	-	0.272 (0.160)
Nonwhite	-	-	-	-0.272 (0.190)
6–12 credential	-	-	0.268[b] (0.137)	0.255 (0.139)
Experience	-	-	0.196*** (0.039)	0.200*** (0.040)
Math PD	-	-	-	0.020 (0.025)
Collabor. (COL)	-	-	-	-0.076 (0.129)
Topics (TOP)	-	-	-	0.120 (0.144)
COL × TOP	-	-	-	0.182 (0.139)
Slope			-	
Intercept	-	-0.066** (.022)	-0.109*** (0.032)	-0.105 * (0.044)
Nonwhite	-	-	-	0.070 (0.058)
Early entry	-	-	-	-0.004 (0.042)
Math PD	-	-	0.013[c] (0.007)	0.010 (0.008)
COL	-	-		0.023 (0.036)
Grade 6, 7 (GR67)	-	-	-	-0.006 (0.040)
COL × GR67	-	-	-	-0.013 (0.040)
Random effects [a]				
Level 2				
Var. initial status	0.351	0.767	0.636	0.626
Var. slope	-	0.0267	0.0258	0.0263
Corr. init. & slope	-	- .43	- .39	- .40
Level 1				
Residual variance	0.341	0.252	0.252	0.252
Goodness-of-fit [a]				
Initial status R^2	-	-	.17	.18
Slope R^2	-	-	.03	.01
$R^2_{y\hat{y}}$	.756	.865	.863	.863
Log-likelihood	-709	-700	-691	-705
AIC	1425	1411	1401	1450
BIC	1438	1438	1445	1538

Note. Standard errors for the regression coefficients are given in parentheses.
* p < .05; ** p < .01; *** p < .001.
[a] The random effects and goodness of fit estimates are the median estimates across the 50 imputed data sets.
[b] The *p*-value for this estimate is .051.
[c] The *p*-value for this estimate is .078.

The $R^2_{y\hat{y}}$ statistic increased from .75 in the unconditional means model (Model 1) to about .86 with the inclusion of time and other predictors in the three other models (see Table 29), indicating a substantial increase in ability of the model to predict the observed data. Although about 17 of variance in initial KE status was explained by predictors in Model 3 and 18 of the this variance was explained by predictors in Model 4, very little—3 or less—of the variance in slope was explained by the predictors in either model. These estimates of explained variance are so small that they are not meaningful different than zero even if they do reflect slight improvement in the ability of Model 3 and 4 to explain the rate of change in KE over that of Model 2.

In Model 3 of KE, the indicator for nonwhite was not statistically significant. Teaching experience was the only statistically significant predictor of initial KE status in Model 3, and the relationship was positive and strong. The indicator for a secondary credential predicting initial KE status was very close to statistical significance at the .05 level. Collaborative activity and experience teaching ratio and proportion were not significantly related to initial KE status. The estimate for the overall slope was statistically significant and negative. This estimate was -0.11 *SD* per month, and a little less than -.4 *SD* over the 4-month period of the study. Collaborative activity was not related to the rate of change in KE either, nor were the indicator for teaching Grades 6 or 7 and the interaction between those variables. The only predictor for rate of change in KE that approached statistical significance was mathematics PD. The estimate was positive, but it had a *p*-value of .078, and thus did not attain significance at the .05 level.

Summary of findings. The research questions for this study were as follows: 3. What features of teachers' experience in schools are associated with change in mathematical proficiency for teaching multiplicative reasoning topics? 4. How does change in mathematical

proficiency for teaching multiplicative reasoning topics differ (a) across schools, (b) between teachers who teach that content and those who do not, and (c) with the frequency of teachers' collaboration and collegial activity focused on student learning? I found evidence of change in the teachers' MKT, PE, and KE over the semester of the study. On average, the teachers' MKT increased, and their PE and KE decreased. Several factors contributed to the teachers' mathematical proficiency for teaching multiplicative reasoning topics at the beginning of the semester as measured by those three outcomes. Teaching experience was positively related to the teachers' initial status on all three outcomes, and the parameters for teaching experience were statistically significantly in the best-fitting models (Model 3 for MKT, PE, and KE). Having a secondary (Grade 6–12) teaching certificate was positively associated with the teachers' MKT and KE, but statistically significant only in the model of MKT. Collaboration and experience teaching multiplicative reasoning topics were related to initial values in teachers' MKT and PE, with interaction terms reaching (PE) and nearly reaching (MKT) statistical significance at the .05 level. The teachers with more frequent collaborative activity had higher PE than those who collaborated less frequently, but only if they were also teaching multiplicative reasoning topics. In contrast, the teachers who were teaching multiplicative reasoning topics had higher initial MKT than those who were not, but only if they were also working with colleagues.

The models of KE explored in this study did not explain the variation in the rate of KE change over the semester, and only one predictor—mathematics professional development—was close to reaching statistical significance ($p = .078$). This predictor was positively associated with slope and can be interpreted to mean that involvement in professional development mitigated the model-predicted decrease in KE over the semester of the study, but with the caveat that there is a

1 in 13 chance of observing an association of similar or greater strength when no such relationship exists.

The models of MKT and PE explained more of the observed variation in rate of change than the models of KE. In the interpreted model of MKT (Model 3), the early entry indicator and the mathematics professional development variable explained 50% of the variance in an individual's rate of MKT change, with early entry positively—and mathematics PD negatively—associated with MKT change. This finding means that half of the observed differences between how the teachers' MKT changed over the semester were explained by whether or not they were early entry and whether or not they were taking PD. The early entry teachers' MKT increased faster than that of those who were not early entry, and (somewhat surprisingly) the MKT of those in mathematics PD increased more slowly than that of those not in mathematics PD. I did not ask about the content of the PD, so those teachers in mathematics PD may have been focused on learning in a different domain of mathematics that did not affect their MKT related to multiplicative reasoning topics.

Model 3 of the PE outcome accounted for 20 of the variance in the individual rate of PE change. Thus, predictors in this model explained about one-fifth of the variation in the rate of change in PE observed across the teachers in the study. The significant predictors of the rate of change included collaborative activity, the indicator for teaching Grades 6 or 7, and the interaction between these variables. The interaction was statistically significant and can be interpreted to mean that the teachers who collaborated more frequently with others had a slower decrease in PE than those who collaborated less frequently, unless they taught in Grades 6 or 7. For Grades 6 and 7 teachers, more frequent collaboration was associated with a faster decrease in PE.

Qualitative Results

In this section, I discuss results from the qualitative data in the context of the results presented in the previous section. I take each research question in turn, and for each unit of analysis (the school and individual) I discuss findings from the retrospective interviews. Then I discuss how these results fit together with those from the quantitative analyses discussed above.

Question 3. What features of teachers' experience in schools are associated with change in mathematical proficiency for teaching multiplicative reasoning topics? I used the individual teacher as the unit of analysis to analyze the participants' accounts of their own teaching practice and how it had changed since they began teaching. All of the interviewed teachers with useable data (n = 17) reported increased confidence in their knowledge of the mathematics they taught (KE) and in their ability to teach mathematics well (PE). Their descriptions of changing practice also reflected increased MKT: greater content and curricular knowledge, better understanding of student misconceptions, more effective explanations of difficult concepts, and increased knowledge of representations (including manipulatives) and how to use them effectively.

To understand how the teachers explained their growth to themselves, I asked them at the end of the interview which of the many things we had discussed—university coursework and field experience, mentoring and coaching, common lesson planning and other forms of collaborative work with other teachers, professional development, trial and error in their own classroom, and curricular materials—had been most important for learning to be a better mathematics teacher. Of the 17 participants, 12 named collaborative work with other teachers as the most important factor in learning to teach mathematics better. For example, one participant said,

> I would definitely say that having that group of teachers at school, working collaboratively with the other sixth-grade math grade teachers who are teaching

> exactly what I'm teaching and to the exact same age group that I'm teaching has been invaluable for me.

Another participant said,

> The biggest thing would be working with my peers, the other math teachers, my mentor math teacher, who is still a mentor to me, even though she's in eighth grade and I'm in seventh, or the group that I'm working with now. That has to be the number one thing … And my course work, I thoroughly enjoyed. Some classes were useful and some not so useful, but it definitely gave me a broader background than I would have had. But as far as—I'm not the teacher I am because of classwork that I had. It's more because of the experiences I had with other math teachers.

The next most common category of response was feedback from students. Six teachers described interactions with students as the motivation or source of their improved practice. One teacher said,

> So, the one thing is the collaboration, and the other one is the kids leaving here thinking, answering the questions I ask them and getting most of them right. I feel like we've done something today, and the next day is the new day, so we'll figure it out. Basically, collaboration with other teachers and feeling like my students get what I'm talking about.

Another said,

> I just find definitely what makes me a better teacher is the student interaction. That to me is the key. When I get the light bolt moment, that makes me want to be a better teacher. … Just that with that "Oh, I get it!" I love that. That to me is the best part of teaching.

I also asked the participants about professional development, and almost all said that it had only a minor role in their development or had not been useful. On the whole, the professional development they described was general not mathematics specific, limited to a single day or afternoon, and focused on new ideas or novel practices rather than on improvement or master of existing skills. Three teachers, however, credited professional development experiences with significant changes in their teaching practice. One remarked,

> Professional developments are just continuation of learning; it's probably [the] number one [thing that has made the difference]. Constantly learning new ways to teach something and not getting stuck in some kind of rote or practice where you've done it for 3 years.

The other two mentioned particular professional development experiences that had had a profound impact on their teaching and classroom management. One was a professional development workshop on using Socratic questioning techniques, and the other was about differentiated instruction. These kinds of professional development may not have effected the survey outcome measures of MKT, PE, and KE.

Question 4. How does change in mathematical proficiency for teaching multiplicative reasoning topics differ (a) across schools, (b) between teachers who teach that content and those who do not, and (c) with the frequency of teachers' collaboration and collegial activity focused on student learning? The retrospective interviews aimed to elicit data from participants on the teachers they worked with—their grade-level departments or mixed-subject teams—and to use those data to make narrative inferences about how knowledge and teaching self-efficacy beliefs in groups of teachers changed over time.

In the end, pursuing this line of analysis was not fruitful because the groups of teachers that participants had worked with and described had little stability. The participants reported frequent changes in schools and grade levels: 6 of the 17 teachers reported working at two or more schools, and all those who had more than 1 year of teaching experience reported working at more than one grade level. Since the teachers in different schools (and even the teachers working at different grade levels in the same school) organized their collaborative work in different ways and to a different extent, most of the participants reported a wide variety of experience with different groups of teachers from year to year. These data offer some elucidation of the absence of variation in mathematical proficiency for teaching and its change at

the school level: The frequent turnover in schools and grade-level teams may undermine the utility of school affiliation for studying teacher change because teachers working at a particular school may have only recently joined that faculty and may be in different roles from year to year.

The data were more useful for understanding change at the individual level. There was strong evidence that individual teachers thought they had experienced change in their knowledge and beliefs. One participant said of her confidence to teach Grade 8 mathematics, "It grows more and more every year." All of the interviewed participants made similar comments about their confidence in knowing the mathematics they were responsible to teach (KE) and about their confidence in being able to teach effectively (PE).

The retrospective interviews highlighted the role of collaboration (Question 4c) and of experience teaching particular topics (Question 4b). Most of the teachers interviewed credited collaboration with other teachers as the biggest factor in improving their own practice, a finding already discussed in the Question 3 subsection.

One advantage of the interview data was the opportunity to learn about the process of improvement and understand the role of collaboration and work with other teachers in this process. Twelve of the 17 interviewees credited work with other teachers as a critical factor in developing their mathematical proficiency for teaching and described a range of activities including planning common lessons and assessments, observing teachers, discussing instructional problems and techniques, and sharing teaching resources. The descriptions frequently included affective descriptions—the participants reported mutual trust; feeling it was "safe" to ask questions; and feeling that their peers were approachable, welcoming, and helpful. One teacher described an experience with a peer when she had been assigned to teach Grade 8 after teaching in her first 2 years at Grade 6:

> One of the eighth grade teachers actually came to me when I was teaching last year and brought me a gift. And it was everything they had done last year on a little flash drive. And she said, "It's going to overwhelm you but you'll have the summer to look through and see what we're going to teach," you know. "It's all in order." And she left her phone number, and she said, "Look it over and start thinking about questions, start thinking about what are you going to need help with and then call me." … I am comfortable going to any one of those math teachers and saying, "I need help with this."

Unlike this teacher, other participants who reported working on common lesson planning or other activities with peers and who did not say that other teachers were important for their professional growth did not describe relationships with a positive affective dimension like the one described in the quotation above. These negative cases provided supporting evidence that a positive affective component in mentoring and collegial relationships may be critical for their utility in helping beginning teachers develop.

The wide range of teachers' assignments made it impractical to ask each teacher about multiplicative reasoning topics in particular, but 8 of the 17 teachers described changes in self-efficacy beliefs for teaching particular topics or at particular grade levels. For example, one participant made distinctions between her teaching self-efficacy for different grade levels, "Now put me in ninth grade and who knows? But sixth grade and seventh grade—so far so good." She also described changes in confidence at an even smaller grain-size than that of grade-level:

> Once I teach something once I feel a lot more confident about it. And some things I walk in knowing exactly how I'm going to teach it. And some things are a little bit trickier. But I guess each new unit is different. I don't really have a pattern that I can say, "Oh, it's gotten easier since last year," or anything like that. I can say that the things I'm teaching this year that I had taught last year already are easier for me.

Fifteen of the teachers made similar distinctions, and six described how being assigned to teach at a different grade level led to an initial drop in confidence in their knowledge of the material and their ability to teach it.

The findings on teachers' assessments of their changing self-efficacy for teaching were surprising given the evidence from the quantitative data that teaching self-efficacy decreased over the semester. The interviews were conducted in February, however, when the teachers' self-efficacy beliefs—according to the survey—were on average higher than they were at the end of the year. One significant factor in understanding these results is the end-of-year exam which was scheduled in mid-April. Scores were not released until the end of June (a month after the conclusion of the study), so the teachers only knew how their students seemed to be doing, not how they had actually performed. Several teachers mentioned receiving the results of the end-of-year exam in previous years and the associated change in their self-confidence. One teacher reported a boost in confidence after receiving the score results:

> The last year I was scared to death that after the CRCT [end-of-year exam] results came back in, that … all the three senior math teachers in the sixth grade you know would have these wonderful scores and mine would be like just miserable. It didn't end up that way, my results were just a little bit behind theirs.

Another teacher reported a similar experience:

> I really expected my scores to not compete with [those of] my colleagues that have been doing this for a long time, but mine were right in the mess with theirs. So when posttest scores came around, which was what we call our finals, they, my scores were right in line with theirs too, so that was a confidence booster. Okay, I am doing something right.

It is possible that these teachers' peers had the same expectations and instead experienced a drop in self-efficacy after seeing that their scores were not better. Such a reaction to disappointing end-of-year scores at a wide scale would explain the drop in self-efficacy beliefs for teaching found in the survey results, and yet still agree with the reports of growing self-efficacy among newer teachers. That teaching experience was not strongly correlated with the rate of change in the outcome variables on the survey may be explained by the diversity of the teachers'

experiences. Only some of the new teachers experienced growth in self-efficacy beliefs for teaching during the semester and other teachers' self-efficacy beliefs decreased.

CHAPTER 5

THE UNITED STATES STUDY

There is a critical lack of large-scale research studies that compare the relative effects of differences in clinical experiences across a wide range of teacher education programs (Clift & Brady, 2005; Ronfeldt & Reininger, 2012; Wilson et al., 2002). The timing, length, and quality of clinical experiences vary widely within and between nations (Cobb, 1999), and this fact at once demonstrates the need for comparative research and the opportunity to conduct it. In the present study, I exploited the variability in U.S. teacher preparation to provide some of the first large-scale, cross-institutional evidence about the effects of student teaching on mathematical proficiency for teaching. The results of this study have implications for teacher education policy in the United States and internationally.

Rather than examining student outcomes directly, I examined teacher outcomes that are predictive of student achievement. Student teaching might not have the same effects across the wide range of content areas and grade levels for which teachers are prepared, so focusing on a single subject and a population of similarly prepared teachers made sense. Mathematics teacher education urgently needs improvement (e.g., National Mathematics Advisory Panel, 2008; Center for Research in Mathematics & Science Education, 2010), and in this study, I focused on the preparation of primary (Grades K–6) teachers to teach mathematics. I designed the study to address the following research question:

5. How are the timing, length, and quality of student teaching related to prospective teachers' mathematical proficiency for teaching the K–6 curriculum?

Data and Methods

This study used data from U.S. prospective teacher and teacher preparation program surveys conducted by the Teacher Education and Development Study in Mathematics (TEDS-M; Brese & Tatto, 2012). To facilitate future international comparison of the results (which are outside the scope of this study), the analytic sample was restricted to teachers in programs identified by TEDS-M as involving concurrent preparation in content, practice, and pedagogy to teach primary (Grades K–6) mathematics. The TEDS-M data set includes scales of knowledge and beliefs, and the U.S. sample is nationally representative of the public institutions that prepare teachers. The large-scale data sets used in prior research have not included scales of pedagogical content knowledge and have been restricted to single school districts (Boyd et al., 2009; Ronfeldt & Reininger, 2012) or states (Goldhaber & Liddle, 2011; Harris & Sass, 2007).

The relevant TEDS-M sample for the United States consisted of 1,119 prospective primary school teachers in 49 concurrent U.S. preparation programs operated by public institutions. The data set included sample weights that allowed the sample to accurately represent the population of 20,548. A substantial portion of the data from some institutional and individual surveys was missing. Table 30 reports the percentage of missing data on outcomes and key predictors. I used standard data imputation techniques to handle the missing data problem (see Chapter 4 for more details). Because of the large number of variables in the data set related to the participants' prior experience with mathematics, their education, and their teacher education programs, I was confident that the imputed values for the outcome measures of teachers knowledge and beliefs were appropriate.

Little is known, however, about how programs for prospective teachers set policies related to the length and timing of student teaching—key predictors of substantive interest in the

present study. As a result, I had little confidence that the data set included reasonable predictors of those variables and was concerned that imputation mechanisms would not improve estimates over chance. In fact, preliminary analyses with the fully imputed data set revealed inflated standard errors for these predictors. Because only 4 programs were missing these data, and because those 4 programs contributed only 75 prospective teachers (7% of the sample), who represented only 1,024 teachers in the population (5% of the sample), I elected to remove those programs and the teachers in them from the analytic sample. Table 30 also reports the percentage of missing data on key predictors and outcome variables for the analytic sample. In most cases, the percentage of missing data was reduced in the analytic sample, indicating that the participants in programs that did not report the timing or length of student teaching were missing data on other variables as well.

Table 30

Percentage of Missing Data on Key Analytic Variables for the Sample, the Analytic Sample, and Those Represented by These Samples

	Key predictor		Outcome	
	Student teaching timing & length	Student teaching quality	Knowledge (both scales)	Beliefs (both scales)
Of sample (N = 1,119)	6.7	24.7	25.5	22.3
Of those represented by sample (N = 20,548)	5.0	25.1	27.5	23.6
Of analytic sample (N =1,044)	0	25.1	25.3	22.5
Of those represented by analytic sample (N = 19,524)	0	25.1	26.8	23.5

I used the R program Amelia II (which employs a bootstrap expectation-maximization algorithm; see Honaker, King, & Blackwell, 2011) to create 50 imputed data sets for all 1,119 individuals in the sample using 70 individual- and program-level variables from the TEDS-M

data set. Thus, the observed data from the 75 individuals with missing data on student teaching length and timing and who were excluded from the analytic sample nevertheless contributed to the imputation of missing data for other individuals. The missing data on the variables used for imputation ranged from 0 to 28%.

Predictors of Substantive Interest

The TEDS-M study distinguished *introductory field experiences* (e.g., observation) from *extended teaching practice*, defining the latter as "two weeks or more of continuous work in schools when the main purpose is to prepare and enable future teachers to assume overall responsibility for teaching a class or classes of students" (Tatto et al., 2012, p. 109). I did not consider introductory field experiences in the present study, and I use the term *student teaching* to refer to extended teaching practice, as defined in the TEDS-M study.

All the U.S. teacher preparation programs in TEDS-M reported some student teaching, but the programs varied with respect to the timing of student teaching, with many reporting student teaching prior to the final year of the program. The student teaching timing variable was a binary indicator variable: A value of 1 was defined as attending a teacher preparation program that involved student teaching prior to the final year of preparation, and a value of 0 was defined as attending a program that did not involve student teaching prior to the final year, a so-called normal student teaching timing. About 45% of the U.S. institutions in the analytic sample (20 of the 44) scheduled student teaching before the final year of the teacher preparation program, and after adjusting for the sample weights, an estimated 37.6% of the prospective teachers represented by this sample experienced early student teaching. Early student teaching timing did not look the same at every institution. In some institutions, early student teaching involved 20 to 30 six-hour days. In other schools, the early student teaching lasted several months or a whole

year but was only 2 to 4 hours a day. The evidence available from the survey data points clearly to one common feature: All programs with early student teaching involved a student teaching experience that was either concurrent with or followed by other preparation activities such as content and methods coursework.

The length of student teaching was operationalized as the total number of student teaching hours taken as the product of days of extended teaching practice per year and the corresponding annual estimate of the average number of hours per day. The unit of weeks—defined as 40 contact hours—was used for analysis to aid interpretation. This point bears stressing because in the following analysis, the unit *week* could refer to 40 contact hours stretched over a whole month at 2 hours per weekday or it could refer to five 8-hour days.

The length of student teaching in the sample institutions ranged from 6 weeks (240 contact hours) to 30.6 weeks (1224 contact hours), but calendar time for these same student teaching experiences ranged from approximately 1 to 3 semesters. Programs that had early student teaching also generally had longer student teaching (see Figure 2). The median length of student teaching was 16 weeks (640 contact hours). To aid interpretation in multilevel regression modeling, it is convenient to center variables so that a score of 0 is within the observed data range and has some substantive meaning. Predictors are frequently centered on the mean, but there were no observed programs with the mean program length. Instead, I subtracted the median (16 weeks) from the student teaching length variable before doing any analysis.

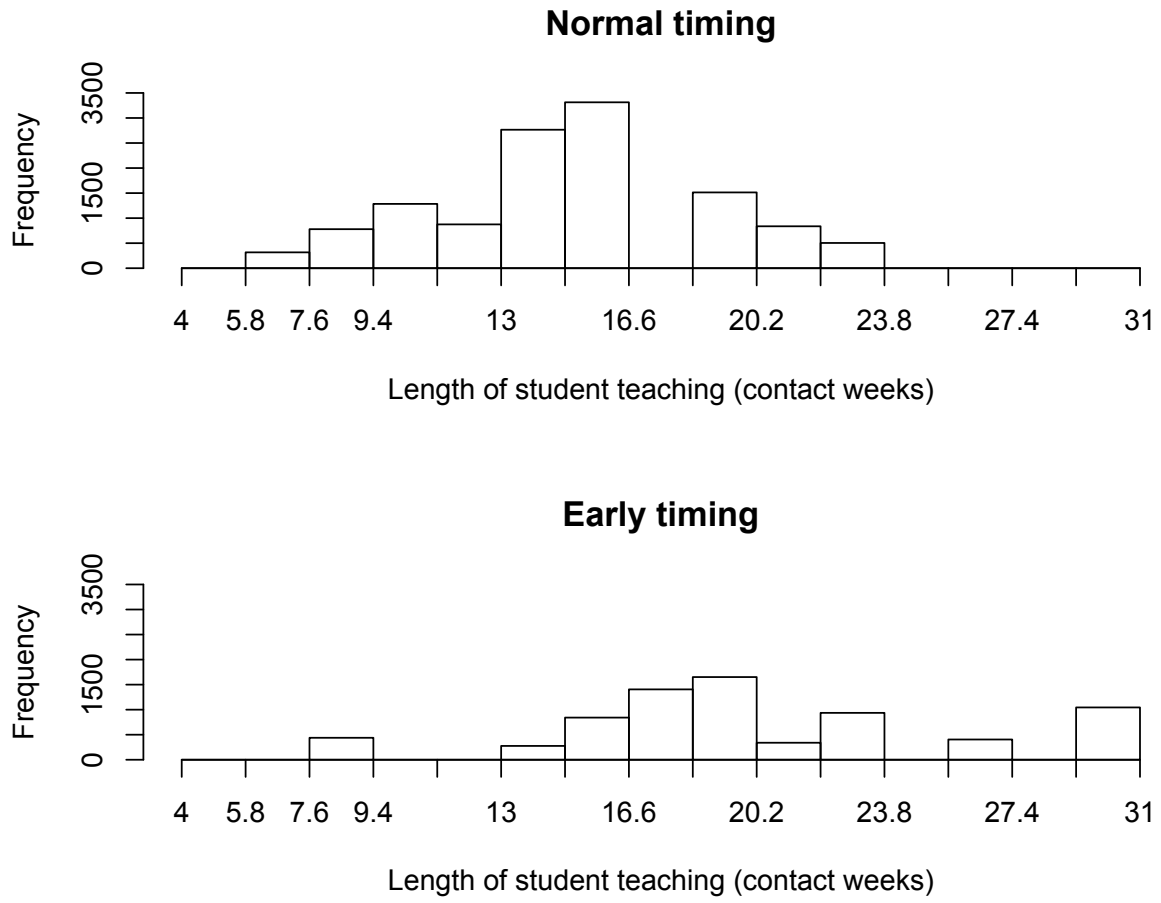


Figure 7. Histograms showing the distribution of student teaching length in programs with normal timing (last year only) and early timing (before the last year) for student teaching; the reported frequencies are population estimates based on sampling weights.

The quality of student teaching was operationalized using a Rasch rating scale model released with the TEDS-M data set (see Table 31; Brese & Tatto, 2012). The resulting continuous variable had a neutral value of 10 and summarized the prospective teachers' rating responses (1: never, 2: rarely, 3: occasionally, 4: often) on questions about the frequency of various student teaching activities that can support connections between teaching and children's learning. Using a Rasch rating scale model enabled the TEDS-M researchers to assess whether the instrument would provide valid comparisons between teachers in different countries; the variable was modeled as a latent rather than manifest variable because each rating item captured

the perception of the respondent rather than an objective measure of frequency. Rather than mean centering the student teaching quality variable before analysis, I subtracted 10 because the scores were designed with a neutral position of 10 points.

Table 31

Quality of Student Teaching Measure: Frequency of Opportunities to Connect Teaching and Learning

Item	During the school experience part of your program, how often were you required to do each of the following? (1: Never, 2: Rarely, 3: Occasionally, 4: Often)
1	Observe models of the teaching strategies you were learning in your courses
2	Practice theories for teaching mathematics that you were learning in your courses
3	Complete assessment tasks that asked you to show how you were applying ideas you were learning in your courses
4	Receive feedback about how well you had implemented teaching strategies you were learning in your courses
5	Collect and analyze evidence about pupil learning as a result of your teaching methods
6	Test out findings from educational research about difficulties pupils have in learning in your courses
7	Develop strategies to reflect upon your professional knowledge
8	Demonstrate that you could apply the teaching methods you were learning in your courses

Note. These items were combined under a Rasch rating scale model to produce a continuous variable with neutral value of 10 named MFB13CLP; see the forthcoming TEDS-M 2008 Technical Report available at http://www.iea.nl/teds-m.html for more information.

Outcome Measures

The four outcomes used in this study were operationalized using IRT scales developed for the TEDS-M project. These outcomes were well matched to the underlying constructs I was interested in studying and well aligned with the research question and framework of the study. Two outcomes were measures of teacher knowledge. The TEDS-M instrument for mathematics pedagogical content knowledge was based explicitly on Shulman's work (1986), and LMT items were used in its development. This measure did not focused on multiplicative reasoning, and

instead assessed PCK for the K–6 mathematics curriculum, including 32 items from the content domains of algebra, geometry, number, and data and the pedagogical domains of planning and enacting (see Brese & Tatto, 2012). The TEDS-M content knowledge instrument assessed teachers' knowledge in the same content domains and included 74 items. Because of the rotated block design and booklet administration of these tests, participants only answered a subset of the items that composed each instrument (Tatto et al., 2012). Both scales had a mean score of 500 and a standard deviation of 100 scale points.

The other two outcomes used in the present study were measures of teachers' beliefs about teaching and learning mathematics. These TEDS-M measures were created using Rasch rating scale models to aggregate participants ratings of agreement across a number of statements that captured two kinds of beliefs about mathematics teaching and learning: beliefs about mathematics as a process of inquiry (abbreviated here as *math-as-inquiry* beliefs) and beliefs about learning mathematics through active involvement (abbreviated here as *active learning* beliefs). The items that comprised these two scales are presented in Table 32. Each scale was constructed with a neutral value of 10. The TEDS-M researchers described these measures as "largely consistent" (Tatto et al., 2012, p. 157) with the work of Philipp (2007) and Staub and Stern (2002).

Covariates to Mitigate Selection Bias

Selection bias is a concern when comparing teachers in different programs; individual characteristics rather than program characteristics may be responsible for the observed outcomes. Individual-level and program-level covariates similar to those used in earlier research (Boyd et al., 2009; Ronfeldt & Reininger, 2012) to mitigate bias in estimates of the effects of student teaching duration and quality were available in the TEDS-M dataset and were used in the present

analysis. These included age, gender, socioeconomic status, length of the program, and (average) student achievement in secondary school at both the program and individual levels.

Table 32

TEDS-M Measures for Teachers' Beliefs: Mathematics-as-Inquiry and Active Learning Beliefs

	Mathematics-as-Inquiry Beliefs [a]
Item	To what extent do you agree or disagree with the following beliefs about the nature of mathematics? (1: Strongly disagree, 2: Disagree, 3: Slightly disagree, 4: Slightly agree, 5: Agree, 6: Strongly agree)
1	In mathematics many things can be discovered and tried out by oneself
2	If you engage in mathematical tasks, you can discover new things (e.g., connections, rules, concepts)
3	Mathematical problems can be solved correctly in many ways
4	Many aspects of mathematics have practical relevance
5	Mathematics helps solve everyday problems and tasks
	Active Learning Beliefs [b]
Item	From your perspective, to what extent would you agree or disagree with each of the following statements about learning mathematics? (1: Strongly disagree, 2: Disagree, 3: Slightly disagree, 4: Slightly agree, 5: Agree, 6: Strongly agree)
1	In addition to getting a right answer in mathematics, it is important to understand why the answer is correct
2	Teachers should allow pupils to figure out their own ways to solve mathematical problems
3	Time used to investigate why a solution to a mathematical problem works is time well spent
4	Pupils can figure out a way to solve mathematical problems without a teacher's help
5	Teachers should encourage pupils to find their own solutions to mathematical problems even if they are inefficient
6	It is helpful for pupils to discuss different ways to solve particular problems

Note. Items were combined under a Rasch rating scale model to produce a continuous variable with neutral value of 10; see the forthcoming TEDS-M 2008 Technical Report available at http://www.iea.nl/teds-m.html for more information.
[a] The TEDS-M derived variable MFD1PROC.
[b] The TEDS-M derived variable MFD2ACTV.

Estimates of the effects of teacher preparation programs can combine selection into the program with program features (Goldhaber & Liddle, 2011; Harris & Sass, 2007), so I also included program-level variables related to selectivity and individual-level high school achievement to control for self-selection bias. The sets of pretreatment covariates at the program

and individual levels also included many of the variables used to predict college choice (e.g., Cabrera & La Nasa, 2000) and teacher knowledge (e.g., Hill, 2007, 2010). Thus, those covariates likely reduce bias from omitted variables. The covariates included the number of mathematics and mathathematics pedagogy classes, the importance of standardized tests for selection into the program, and circumstances hindering program participation.

Descriptions of the covariates used are presented in Table 33. In addition to the covariates already described, I included two scales related to student teaching and one scale related to the program quality overall. The Student Teaching Supervisor Reinforced University Goals scale included items such as, "I learned the same criteria or standards for good teaching in my courses and in my student teaching." The Student Teaching Supervisor Feedback Quality scale included items such as, "The feedback I received from my supervising teacher helped me to improve my understanding of pupils." And the Coherence of Preparation Program scale included items such as, "Later courses in the program built on what was taught in earlier courses in the program." These items were rated on a 4-point scale from *disagree* to *agree*. I also included as individual level covariates three scales of the tertiary mathematics topics prospective teachers had studied. The scales were designed for use with secondary teachers as well, and many of the topics are not frequently studied by Grades K–6 prospective teachers in the United States. Still the scales provided a way of distinguishing different levels of post-secondary mathematical training among the prospective teachers in the sample. An example of a Continuity scale topic was limits; an example of a Discrete structures scale topic was prime numbers; and an example of a Geometry scale topic was Euclidean axioms. The topics used to form these scales are described in more detail in Appendix D.

Table 33

Covariates Used to Mitigate Selection Bias.

Covariate	Scale	Question or description
Program level		
Length of program	Number of years	Length of the program was reported in years and months; this is the sum as a value in years.
Math and math pedagogy classes	Number of classes	Sum of the reported number of classes for mathematics pedagogy, mathematics content related to the school mathematics curriculum, and academic mathematics
Average student achievement	6-point scale	With reference to national achievement norms for their age group, are students? 1: Far below average, 2: Below average, 3: Average, 4: Above average, 5: High (top 20%), 6: Very high (top 10%).
Importance of standardized tests for selection	4-point scale	How important is candidates' performance as measured by their performance on a national or state examination in selecting prospective teachers entering the program? (1: Not considered, 2: Not very important, 3: Somewhat important, 4: Very important)
Individual level		
Student teaching supervisor reinforced university goals	Rasch rating scale, neutral is 10 points	Rated 5 statements on a 4-point scale (1: Disagree to 4: Agree), e.g., "I learned the same criteria or standards for good teaching in my courses and in my student teaching;" TEDS-M variable MFB14STR[a]
Student teaching supervisor feedback quality	Rasch rating scale, neutral is 10 points	Rated 4 statements on a 4-point scale (1: Disagree to 4: Agree), e.g., "The feedback I received from my supervising teacher helped me improve my teaching methods;" TEDS-M variable MFB14STF [a]
Secondary school achievement	5-point scale	What was your level of grades compared with your class? 1: Below average, 2: About average, 3: Above average, 4: Near the top, 5: Always at the top.
Age	Number of years	Participants reported their age in a blank space in response to the question, How old are you?
Gender	Binary indicator	Participants reported their gender by checking Female or Male in response to the question, What is your gender?
Socioeconomic status	Raw score ranging from 3 to 19	Sum of ratings on the following scales: (i) 5-point scale on the number of books, (ii) 7-point scale on mother's education, and (iii) 7-point scale on father's education.
Hindering circumstances	Raw score ranging from 3 to 6	Sum of ratings (1: No, 2: Yes) on following circumstances hindering studies: (i) had family responsibilities that made it difficult to do my best; (ii) had to borrow money; (iii) had to work a job.
Coherence of preparation program	Rasch rating scale, neutral is 10 points	Rated 6 statements on a 4-point scale (1: Disagree to 4: Agree), e.g., "Later courses in the program built on what was taught in earlier courses in the program;" TEDS-M variable MFB15COH [a]
Tertiary mathematics		Prospective teachers checked each topic as "Studied" or "Not studied."
Continuity & functions	Rasch scale score, neutral is 10 points	Based on 5 topics, e.g., "Beginning Calculus Topics (e.g., limits, series, sequences);" TEDS-M variable MFB1CONT [a]
Discrete structures & logic	Rasch scale score, neutral is 10 points	Based on 6 topics, e.g., "Number Theory (e.g., divisibility, prime numbers);" TEDS-M variable MFB1DISC [a]
Geometry	Rasch scale score, neutral is 10 points	Based on 4 topics, e.g., "Axiomatic Geometry (e.g., Euclidean axioms);" TEDS-M variable MFB1GEOM [a]

[a] See the forthcoming TEDS-M 2008 Technical Report for more information on these variables (http://www.iea.nl/teds-m.html).

I elected to use covariates to minimize bias rather than other methods, such as propensity score matching techniques. Building a valid propensity estimation model was not feasible given the current state of the literature on the selection of individuals into teacher education programs and on the determinants of program policies for student teaching.

Multilevel Modeling With Survey Weights

The statistical software MPLUS (Version 6.11 for Mac) was used to estimate a separate multilevel model for each outcome variable (prospective teachers nested within preparation programs). The software also estimates standard errors of individual regression coefficients and the likelihood statistic for testing nested models. The complex sampling design of the TEDS-M data was addressed by incorporating sampling weights into the analysis. Weights should not be used without appropriate scaling because unscaled weights can bias estimates (Carle, 2009). Both scaling methods recommended by Carle (2009; cluster sample size and effective cluster sample size) were available in MPLUS, and I used both methods and compared the results. I also ran the analyses without weights. The results across all three methods were very consistent with each other.

The statistical model for this study (Equation Set 1) was adapted from VanderWeele (2008) and is appropriate for estimating neighborhood effects—effects at the program rather than individual level. This model accommodates the expected homogeneity among prospective teachers in the same program (Gelman & Hill, 2007).

$$
\begin{aligned}
Y_{ij} &= \mu_j + \gamma_1 Q_{ij} + \beta_1 X_{ij} + e_{ij} \\
\mu_j &= \alpha + \gamma_2 T_j + \gamma_3 L_j + \gamma_4 T_j L_j + \beta_2 Z_j + u_j \quad (1) \\
e_{ij} &\sim N(0,\sigma_1);\ u_j \sim N(0,\sigma_2)
\end{aligned}
$$

The first equation expresses the individual level of the model. The model predicts the outcome Y (prospective teachers' knowledge or beliefs) with i indexing individuals and j

indexing programs. The matrix Q_{ij} represents the student teaching quality experienced by person i in program j; the matrix X represents the individual-level covariates (for a detailed listing, see Table 33); and e_{ij} is the random error term associated with individual i in program j. The coefficient γ_1 and vector of coefficients β_1 are estimated by fitting the model to the observed data; these terms provide estimates of the relationships between these variables and the outcome Y. The last term in the first equation is the intercept term μ_j. It represents the average outcome for each program; predicting this term using program-level variables is the role of the second equation.

The second equation includes T_j, the binary indicator variable representing whether program j has early student teaching. It also includes L_j, the variable representing the length of the student teaching in program j. The matrix Z represents the program-level covariates (for a detailed listing, see Table 33). The interaction term T_jL_j expresses the possibility of an increase in the outcome for timing and length beyond that accounted for by each independently. Finally, u_{ij} is the random error term associated with program j. A fitted model provides estimates of the coefficients γ_2 and γ_3 and of the vector of coefficients β_2. I also estimated the coefficient for the program-level interaction between timing and length, γ_4. The last line of Equation Set 1 indicates the assumption that the error terms are normally distributed.

Results

The research question for this study asked: What are the effects of earlier, longer, or better student teaching on prospective teachers' knowledge and beliefs with respect to mathematics? To address that question, I modeled teachers' knowledge (content knowledge and pedagogical content knowledge) and beliefs (mathematics as inquiry and active learning) using program- and individual-level covariates to control for selection bias and three substantive

predictors: an indicator for whether the student teaching was timed before the last year, the length of student teaching in weeks (each week defined as 40 contact hours), and the quality of the student teaching operationalized as the frequency of opportunities to connect students' learning to teaching practice. I also included an interaction term between student teaching timing and length at the program level. In this section, I first describe the model selection process and then interpret results from the final models.

Model Fit and Selection

For each outcome, I fit a series of three nested models: the unconditional means or null model for baseline comparison, a model including all covariates, and a full model with the covariates and the substantive predictors. Because these models were nested, consecutive models could be tested using log-likelihood ratio tests (LRT) to determine at each step whether adding more predictors improved model fit. In addition to considering LRT, I also looked at the Bayesian information criterion (BIC) and Akaike's information criterion (AIC). These model criteria are used to compare the fit of nonnested models (particularly when LRTs are not an option) and provide alternative evidence of model fit. For both, the model in a set of putative models that has the smallest criterion is preferred (e.g., Singer & Willet, 2003, pp. 120–122). Table 34 summarizes the model fit results.

For all four teacher outcomes, the LRT showed that the full model (that with the substantive predictors) had significantly improved fit over the model with just the covariates (see Table 34). The information criteria provided consistent information, except in the case of the model for pedagogical content knowledge, in which the smallest BIC selected the model with covariates rather than the full model.

Table 34

Fit Information for the Models of Teacher Outcomes

Model	LRT	*df*	BIC	AIC
1: Pedagogical content knowledge				
a: Null model	-	-	11706.7	11691.9
b: With covariates	122.8 ***	15	11688.2 [a]	11599.1
c: With all predictors	12.1 *	4	11704.0	11595.1 [a]
2: Content knowledge				
a: Null model	-	-	11699.4	11684.6
b: With covariates	143.0 ***	15	11660.7	11571.6
c: With all predictors	38.0 ***	4	11650.5 [a]	11541.6 [a]
3: Beliefs – math as inquiry				
a: Null model	-	-	3882.1	3867.2
b: With covariates	145.0 ***	15	3841.3	3752.2
c: With all predictors	50.7 ***	4	3818.4 [a]	3709.5 [a]
4: Beliefs – active learning				
a: Null model	-	-	3493.9	3479.1
b: With covariates	105.3 ***	15	3492.9	3403.8
c: With all predictors	33.8 ***	4	3487.0 [a]	3378.0 [a]

* $p < .05$; *** $p < .001$.

[a] The model with the smallest Bayesian information criterion (BIC) or Akaike's information criterion (AIC) is y preferred.

In multilevel models (and in multiple regression more generally), it is possible for the model fit to improve with a group of predictors even though none of the predictors is significantly related to the outcomes or meaningful for explanation. In strict prediction models, when the purpose of the model is to identify the most accurate values for an outcome on unobserved values of predictors, the best-fitting model will give the best predictions even in those cases. In the present study, the purpose of the model was dual: The covariates served to predict the outcomes across systematic differences in programs and individuals, and the substantive predictors served an explanatory role. Thus model fit was not sufficient to justify selecting the full model for each outcome. I also considered to what extent each model

explained the individual- and program-level outcome variance (i.e., observed outcome differences within and between programs).

Table 35 reports the partitioned individual- and program-level outcome variance along with the intraclass correlation (ICC) coefficient for each model. The ICC describes the portion of overall variance that exists at the program level. It can also be interpreted as the degree to which the outcomes of individuals in the same program resemble each other. With each consecutive model, the portion of variance explained at the individual and program level increased—suggesting that the covariates were functioning as intended to control for individual and program differences. In addition, the ICCs exhibit a decreasing pattern showing that relatively more of the program-level than individual-level variance was explained by models with more variables. Across outcomes, the covariates and predictors explained 10–15% of the individual-level variance in outcomes and 50–70% of the program-level variance.

Table 35 also reports variance standard errors and significance tests that variance is greater than zero. Of note is the fact that the program-level variance in Model 1b was not significantly different from zero ($p = .124$). This finding suggests that the covariates in Model 1b accounted for the variance in pedagogical content knowledge at the program level, and that very little variance remained at the program level to be accounted for by the substantive predictors in Model 1c. This means that the program-level differences in prospective Grades K–6 teachers' PCK cannot be explained with student teaching variables. By contrast, content knowledge variance at the program level remained significantly different from zero even under the full model. This result calls into question the utility of Model 1c in explaining prospective Grades K–6 teachers' PCK.

Table 35

Partitioned Level 1 (Individual) and Level 2 (Program) Variance in Teacher Outcomes

	Level 1		Level 2		
Model	Variance (SE)	Variance explained	Variance (SE)	Variance explained	ICC
1: Pedagogical content knowledge					
a: Null model	4074.0 *** (251.5)	-	330.8 * (142.5)	-	.067
b: With covariates	3699.2 *** (222.4)	.092	114.8 (74.7)	.653	.032
c: With all predictors	3670.9 *** (217.8)	.099	89.2 (63.9)	.730	.036
2: Content knowledge					
a: Null model	3915.5 *** (241.3)	-	880.0 ** (254.3)		.189
b: With covariates	3481.5 *** (220.4)	.111	436.7 ** (133.5)	.504	.141
c: With all predictors	3379.1 *** (211.2)	.137	341.1 ** (116.3)	.612	.143
3: Beliefs – math as inquiry					
a: Null model	2.267 *** (0.113)	-	0.173 ** (0.059)	-	.059
b: With covariates	1.992 *** (0.098)	.121	0.103 * (0.043)	.405	.051
c: With all predictors	1.908 *** (0.094)	.158	0.077 (0.039)	.555	.050
4: Beliefs – active learning					
a: Null model	1.555 *** (0.101)	-	0.144 *** (0.041)	-	.109
b: With covariates	1.428 *** (0.087)	.082	0.070 * (0.028)	.514	.094
c: With all predictors	1.392 *** (0.083)	.105	0.049 * (0.024)	.660	.088

* $p < .05$; ** $p < .01$; *** $p < .001$.

The model-fit results suggest five models for final analysis. For the pedagogical content knowledge outcome, conflicting fit results and disturbing information about the variance explained warranted a close comparison of Model 1b and Model 1c. For the other three outcomes, the most detailed models (Models 2c, 3c, & 4c) had the best fit and also exhibited appropriate variance characteristics. Moreover, the results clearly indicated that the substantive predictors that were the focus of this study—student teaching timing, length, and quality—contributed significantly to an explanation of these teacher outcomes. The five models are discussed in detail in the next section.

Final Model Parameters for Knowledge Outcomes

I begin with a comparison of the two models for pedagogical content knowledge and the final model of content knowledge. Both of these outcomes were on a similar scale, which facilitated comparison of the unstandardized regression coefficients. First, I evaluated the covariates to check for significance and usual signs. Because the critical role of these variables was to mitigate selection bias (a predictive rather than explanatory role), the statistical significance of individual covariates was not an appropriate criterion for exclusion; even nonsignificant covariates reduce bias. Inappropriate signs, however, especially for statistically significant coefficients, could indicate that the covariates were not functioning as intended.

Table 36 shows that average student achievement was the only statistically significant program-level covariate for Models 1b, 1c, and 2c; it had a positive sign, as expected. On the individual level covariates, secondary school achievement and a self-report measure of preparation program coherence were significant covariates and positive across all three models. Although not significant, variables measuring SES and circumstances hindering students' participation in the teacher education program had the expected signs. Student teaching supervisor covariates were nonsignificant except for the coefficient of supervisor feedback on pedagogical content knowledge. Rather than an implausible causal effect (feedback that helps one learn reduces one's knowledge), this result could be a signal that supervisors were systematically spending more time working with those prospective teachers with relatively low knowledge. Because the data set did not include preprogram knowledge measures, these students could have shown lower final outcomes even if they had made significant gains in knowledge during their student teaching.

Table 36

Predictors of Teachers' Pedagogical Content Knowledge and Content Knowledge

Predictor	Model 1b: covariates *B (SE)*	Model 1c: all predictors *B (SE)*	Model 2c: all predictors *B (SE)*
Intercept	551.8 *** (6.35)	551.4 *** (7.18)	530.6 *** (7.54)
Z – Program level covariates			
Length of program (years)	3.63 (4.35)	2.14 (3.84)	0.42 (5.51)
Number of math & math pedagogy classes	-1.04 (1.49)	-0.76 (1.41)	0.78 (2.11)
Avg. student achievement	6.83 * (3.44)	6.80 * (3.39)	12.13 ** (4.31)
Importance of standardized tests for selection	0.70 (2.27)	1.33 (2.17)	3.18 (2.78)
X – Individual level covariates			
Student teaching supervisor reinforced university goals	-0.45 (1.65)	0.16 (1.66)	-1.85 (1.62)
Student teaching supervisor feedback helped improve teaching	-2.80 * (1.34)	-2.50 (1.35)	-0.68 (1.41)
Secondary school achievement	11.53 *** (2.53)	11.52 *** (2.47)	14.45 *** (2.09)
Age (years)	-0.70 * (0.35)	-0.77 * (0.35)	-0.32 (0.40)
Gender (male)	5.88 (7.97)	5.93 (7.95)	22.68 ** (8.10)
SES (e.g., mothers' education)	2.24 * (1.10)	2.07 (1.10)	1.83 (0.97)
Hindering circumstances (e.g., need to work)	-2.70 (3.12)	-2.99 (3.09)	-1.16 (2.68)
Coherence of preparation program	3.74 *** (1.09)	4.27 *** (1.15)	3.95 ** (1.20)
Tertiary math topics studied			
Continuity & functions	6.21 ** (2.06)	6.16 ** (2.04)	7.77 *** (2.11)
Discrete structures & logic	-4.63 ** (1.59)	-4.42 ** (1.57)	0.20 (1.63)
Geometry	-3.42 (1.83)	-3.38 (1.81)	-7.35 *** (1.90)
Student teaching variables			
T – Early timing (0 or 1)		8.76 (7.19)	15.37 (10.55)
L – Length (40-hr weeks)		0.99 (0.95)	2.52 ** (0.96)
T × **L** – Timing & length interaction		-0.86 (1.20)	-3.92 * (1.58)
Q – Freq. of opportunities to connect teaching and learning		-3.14 (1.76)	-7.23 *** (1.45)

* $p < .05$; ** $p < .01$; *** $p < .001$.

There were several differences between the models for pedagogical content knowledge and content knowledge. Age was a significant (and negative) predictor for pedagogical content knowledge but not a significant predictor for content knowledge. Gender was positive for both

outcomes, but significant only for common content knowledge—indicating a significant gender gap in the content knowledge of prospective Grades K–6 teachers in this sample. There were 111 men (about 10%) in the sample.

The tertiary mathematics topics covariates also require some discussion. Continuity topics (e.g., those encountered in a class on series or calculus) were significantly and positively related to both knowledge outcomes, whereas discrete structures (e.g., linear algebra and number theory) and geometry (e.g., Euclidean, analytic, or differential geometry) were significantly and negatively related to knowledge outcomes. Regression coefficients are generally interpreted by holding all other variables fixed—which in this case would lead to the somewhat implausible interpretation that studying geometry leads to less content knowledge.

In fact, the number of tertiary mathematics courses taken by prospective elementary teachers is likely restricted. Increasing the number of continuity topics studied would almost certainly decrease the number of geometry (or discrete structure) topics studied. Seen this way, the coefficients together represent a trade-off in teacher knowledge: One might expect the prospective teachers who studied more continuity topics (and thus fewer geometry or discrete structure topics) to have had a relative advantage of about 10 to 14 points on the knowledge outcomes compared with those prospective teachers who studied more geometry or discrete structure topics.

It is also entirely possible that these results do not indicate a causal mechanism: Those prospective elementary school teachers who took calculus might have done better on the knowledge measures than those who did not take calculus because of a common cause—their experiences studying mathematics in secondary school, for example. In addition, these results might simply reflect the content focus of the teacher knowledge measures: If the instruments had

more items that involved continuity topics than geometric topics, then one would expect those participants with more tertiary exposure to those ideas to do better on the instrument than those with less experience. I examined the framework for the content knowledge instrument (Tatto et al., 2012, pp. 129–132) and the TEDS-M User Guide (Brese & Tatto, 2012) but was not able to determine the relative weight placed on each subdomain (i.e., number and operations, geometry and measurement, algebra and functions, data and chance) in terms of the fraction of items on the instrument. Overall, the covariates predicting knowledge outcomes appeared to have functioned as expected to mitigate selection bias. I next discuss the substantive predictors—the timing, length, and quality of student teaching.

None of the substantive predictors in Model 1c of pedagogical content knowledge were significant predictors, although the signs for all four (student teaching timing, length, timing-length interaction, and quality) were the same as in the model of teacher content knowledge. Individually, the coefficients for those predictors were not significantly different from 0, although the set provides significantly more information for predicting the teachers' pedagogical content knowledge than a model without these predictors. One mathematical reason for this finding was that little of the variance of pedagogical content knowledge was at the program level, which means that (after controlling for covariates) the programs achieved similar outcomes with respect to pedagogical content knowledge, regardless of the features of student teaching. This result is surprising because of the theoretical reasons that the pedagogical content knowledge—even more than the content knowledge—might have been learned in the context of teaching practice. A possible explanation is that the preparation programs did little to influence the prospective teachers' pedagogical content knowledge because the classes and student teaching experience had not been designed for that outcome. Given the widely cited research

about the deficits in elementary teachers mathematical knowledge (e.g., Ma, 1999) and the large number of other content areas that generalists must study, mathematics teacher educators may feel they need to spend the available time on content knowledge rather than on pedagogical content knowledge.

The results of Model 2c (content knowledge) pertaining to student teaching timing, length, timing-length interaction, and quality are more interesting. Perhaps most surprising was the result that the student teaching quality coefficient was negatively related to the prospective teachers' content knowledge and that this effect was statistically significant. There are (at least) two possible explanations for this result but neither is entirely satisfactory. First, the result could represent the operation of a causal mechanism—opportunities to connect student learning and practice somehow disrupted what the prospective teachers knew about the content, which led to lower performance on the outcome measures. I find this explanation rather implausible. A second explanation is that the student teaching quality measure was indicative of support for those prospective teachers with weak content knowledge. Under this hypothesis, the weak teachers (hence the negative sign) reported more frequent opportunities to connect their learning with their teaching, because the teacher educators and supervisors in their program had identified those prospective teachers as needing more support than the others. This second explanation is more plausible, but to rule out either would require data that are not available in the TEDS-M data set.

The second result is that the interaction of timing and length was significant. This interaction can be interpreted to mean that the length of the student teaching moderated the effect of its early timing on the teachers' content knowledge. Using the simple slope method (e.g., Bauer & Curran, 2005; Preacher, Curran, & Bauer, 2006), I calculated the region of significance

($\alpha = .05$) for the interaction. From 16.4 to 26.7 weeks, the effect of early timing for student teaching was not significantly different from zero, but when student teaching length was less than 16.4 weeks or more than 26.7 weeks, early timing had a positive effect on the content knowledge outcome that was statistically significant. Very little data are in the upper region of significance, which includes teachers from only two programs, both of which had early timing. The important point to be learned from the upper region of significance is that those teachers in programs with early student teaching that were also very long nevertheless could have had relatively lower content knowledge outcomes after controlling for individual and program-level covariates. This result may suggest a principle of diminishing returns and higher cost for extending student teaching: Programs with very long student teaching (more than 26 weeks) may not have enough university coursework. It is also possible that those programs differ from the other programs in the study in ways that were not accounted for by the covariates included in the estimation model; they may recruit a particular kind of prospective teacher such as military veterans, for example. The available data preclude further inquiry into the nature of these programs.

The lower region of significance is more illuminating because that region included 65% of the sample and (after adjusting for sampling probability) represented more than 50% of the population of publically prepared prospective Grades K–6 elementary school teachers in the United States. The effect of timing was statistically significant below 16.4 weeks, and the effect ranged between 30.2 points at 5 weeks (2 *SD* below the median) and 15.3 points at 16 weeks (the median). Thus timing had an average effect size of approximately .20 in the lower region of significance. Moreover, for student teaching scheduled in just the last year of the program, the effect of each week of student teaching on content knowledge was significant ($p < .01$). Under

Model 2c, a 5.23-week (1 *SD*) increase in student teaching length would have had an estimated effect of 13.18 points on the content knowledge outcome, an effect size of approximately .13 (see Figure 8).

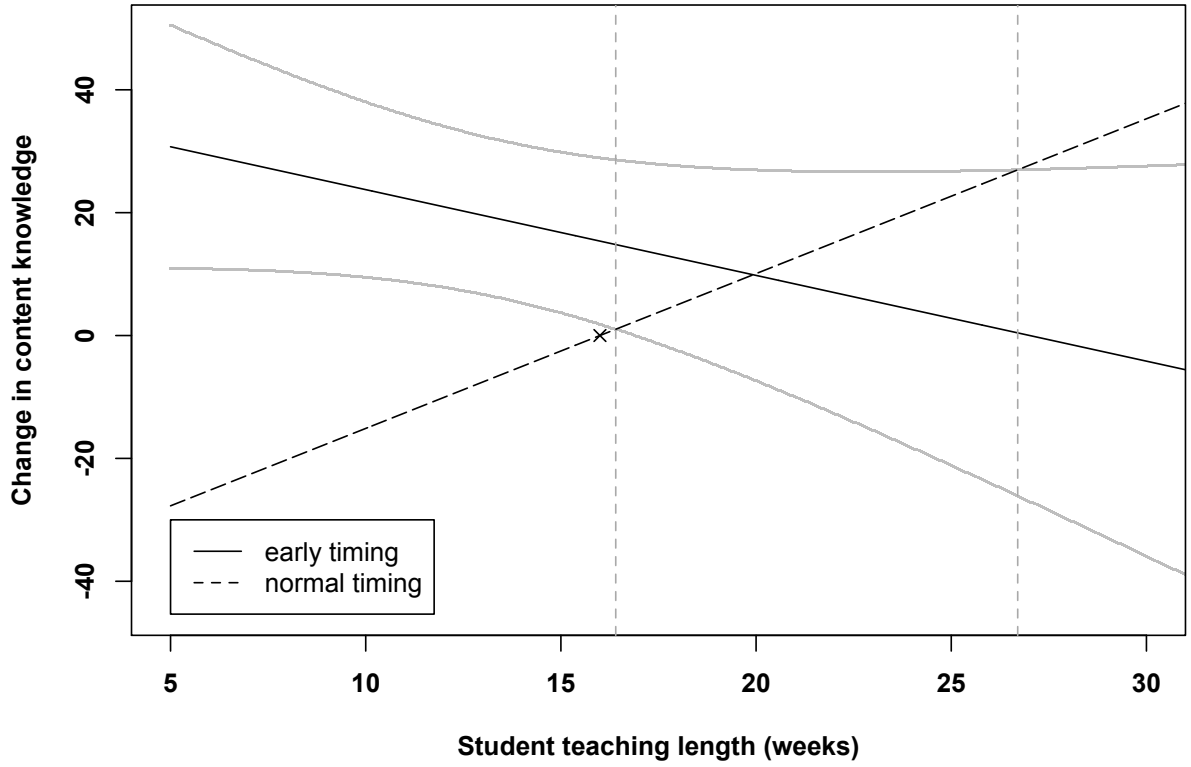


Figure 8. The effect of early timing student teaching on content knowledge with 95% confidence bands as compared to predicted knowledge after 16 weeks of normal timing student teaching (marked ×); early timing had a statistically significant effect when student teaching was less than 16.4 weeks or more than 26.7 weeks.

Final Model Parameters for Belief Outcomes

The results for the belief measures were similar in many respects to the results for the content knowledge outcomes. As with the discussion of the knowledge outcome results, I begin with a summary of the covariates and conclude by reporting results pertaining to the predictors of substantive interest. Because the belief scores ranged from -1.81 to 5.48 with a standard deviation of 1.6, the unstandardized coefficients of predictors in these models were much smaller than those for the same variables in the models of teacher knowledge.

Table 37 shows that of the program-level predictors, only one—the importance of test scores as a selection criterion—was significant in predicting beliefs about mathematics as inquiry. Similarly, only one program level predictor—the length of program—was a significant predictor of active learning beliefs ($p < .05$). It is interesting that the coefficient for this predictor had a negative sign; the sign for coefficients of the same variable in the knowledge models was positive. This negative sign may indicate that the longer programs were less successful than the shorter programs at fostering active learning beliefs.

Many of the individual level predictors were statistically significant for both Models 3c and 4c. The signs of their coefficients were all in the expected direction with the exception of the student teaching supervisor variable on reinforcing university goals. I had expected this predictor to have a positive sign because I assumed that many universities were promoting inquiry-based instruction in pedagogy classes and that a greater reinforcement of university goals would lead to more opportunities for inquiry-based teaching during student teaching. Instead, the coefficient for this variable was negative. This result might have occurred because those schools that lacked inquiry learning goals might also have been the schools that employed supervisors who were supportive of the university's (noninquiry focused) goals and standards. Unfortunately, the data to evaluate this hypothesis do not exist in the US-TEDS-M data set. As with the knowledge models, the coherence of the preparation program had a significant ($p < .001$) positive relationship with the belief outcomes.

Also interesting is the result that the tertiary mathematics topics were not a significant predictor of the beliefs outcomes concerning the teaching and learning mathematics. This result could mean that studying tertiary mathematics did not contribute to the formation of the teachers' beliefs, but it could also simply indicate that the content areas did not contribute more than other

content areas to the teachers' beliefs. I next discuss the predictors of substantive interest: the timing, length, and quality of student teaching.

Table 37

Predictors of Teachers' Beliefs

	Model 3c: math as inquiry *B (SE)*	Model 4c: active learning *B (SE)*
Intercept	1.789 *** (0.164)	1.725 *** (0.141)
Z – Program level covariates		
Length of program (years)	-0.088 (0.083)	-0.170 * (0.072)
Number of math & math pedagogy classes	0.030 (0.035)	0.041 (0.032)
Avg. student achievement	0.104 (0.058)	0.092 (0.054)
Importance of standardized tests for selection	0.095 * (0.048)	0.070 (0.052)
X – Individual level covariates		
Student teaching supervisor reinforced university goals	-0.078 * (0.035)	-0.010 (0.033)
Student teaching supervisor feedback helped improve teaching	0.044 (0.027)	-0.009 (0.021)
Secondary school achievement	0.104 * (0.047)	0.135 ** (0.045)
Age (years)	0.035 *** (0.010)	0.004 (0.007)
Gender (male)	0.531 ** (0.196)	0.345 (0.190)
SES (e.g., mothers' education)	0.032 (0.026)	0.056 * (0.022)
Hindering circumstances (e.g., need to work)	-0.151 * (0.065)	-0.195 *** (0.057)
Coherence of preparation program	0.143 *** (0.028)	0.077 *** (0.022)
Tertiary math topics studied		
Continuity & functions	0.008 (0.052)	0.055 (0.038)
Discrete structures & logic	-0.050 (0.035)	-0.051 (0.035)
Geometry	0.014 (0.040)	-0.064 (0.036)
Student teaching variables		
T – Early timing (0 or 1)	0.265 (0.138)	0.182 (0.111)
L – Length (40-hr weeks)	0.028 (0.017)	0.034 (0.020)
T × **L** – Timing & length interaction	-0.061 * (0.025)	-0.049 * (0.024)
Q – Freq. of opportunities to connect teaching and learning	0.210 *** (0.045)	0.139 *** (0.033)

* $p < .05$; ** $p < .01$; *** $p < .001$.

The first result of interest concerns the quality of student teaching. This effect was highly significant for both models ($p_{3c} < .001$; $p_{4c} < .001$). Moreover, the direction of the effect was positive (unlike the negative estimated effect of quality on content knowledge in Model 2c). For a one standard deviation in frequency of opportunities to connect learning and teaching, the model for math-as-inquiry beliefs predicted an increase in 0.21 scale points with an effect size of approximately 0.13. The model for active learning beliefs predicted an increase in 0.14 scale points with an effect size of approximately .09. These effect sizes are small, yet even so they provide evidence that student teaching can have a role in shifting (if not dramatically shaping) prospective teachers' beliefs about teaching and learning mathematics.

The second result is that the interaction of timing and length was significant for both beliefs outcomes. This result can be interpreted to mean that the length of student teaching moderates the effect of early timing on teachers' beliefs. Using the simple slope method, I calculated the region of significance ($\alpha = .05$) for each interaction. For math-as-inquiry beliefs, the effect of early student teaching was statistically significant and positive when student teaching length was less than 15.8 weeks. For active learning beliefs, the effect of early student teaching was also statistically significant and positive when student teaching length was less than 14.6 weeks. These regions include 42% and 38% of the sample, respectively, and represent 40% and 35% of the population, respectively, of prospective publically prepared Grades K–6 elementary school teachers in the United States.

The effects of an early student teaching on both belief outcomes were positive in the regions of significance, but the regression coefficient for practicum length was not significantly different from zero in either model (see Figures 9 and 10). The statistically significant and positive effect of an early student teaching on math-as-inquiry beliefs ranged between 0.61 scale

points at 5 weeks (2 *SD* below the median) and 0.35 scale points at 13.5 weeks (1/2 *SD* below the median). The standard deviation of the math-as-inquiry outcome was 1.56 scale points, and so early student teaching had an average effect on the math-as-inquiry beliefs of approximately 0.30 *SD* in the region of significance. Similarly, the effect of an early student teaching on the prospective teachers' active learning beliefs ranged between 0.34 scale points at 5 weeks (2 *SD* below the median) and 0.22 scale points at 13.5 weeks (1/2 *SD* below the median). The standard deviation of the active learning beliefs outcome was 1.31 scale points, and so early student teaching had an average effect on active learning beliefs of approximately .20 *SD* in the region of significance. In neither model was the coefficient for the length of student teaching statistically significant, suggesting that the length of a normally timed student teaching did not significantly affect prospective teachers' beliefs about teaching and learning mathematics.

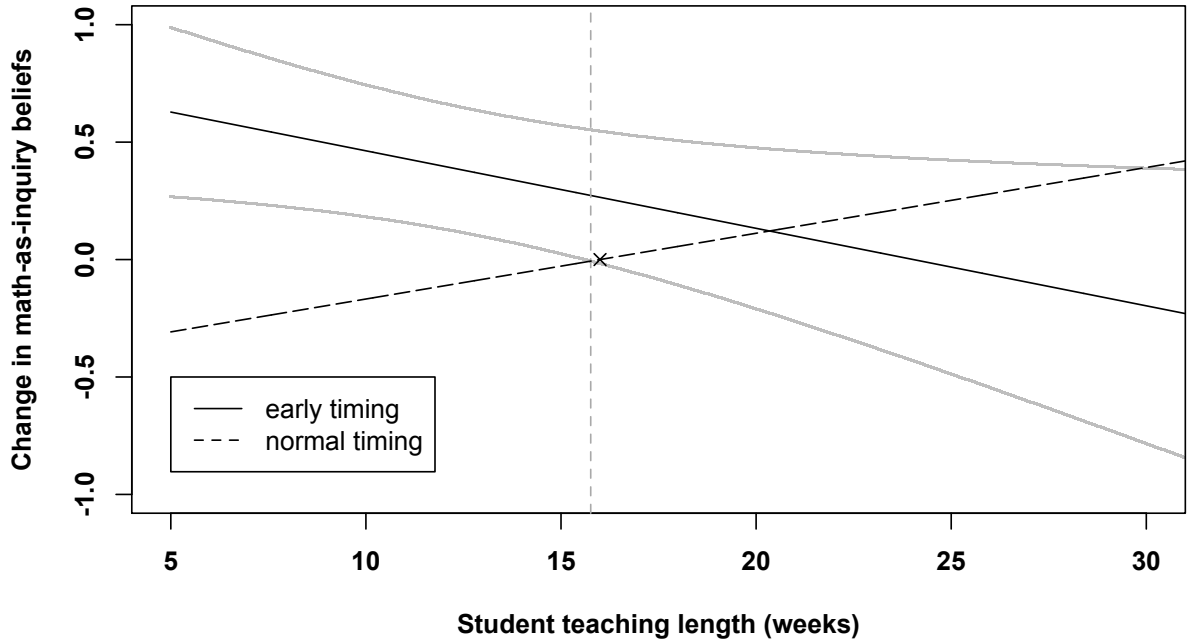


Figure 9. The effect of early timing of student teaching on math-as-inquiry beliefs with 95% confidence bands as compared with predicted beliefs after 16 weeks of normally timed student teaching (marked ×); early timing had a statistically significant positive effect when student teaching was less than 15.8 weeks.

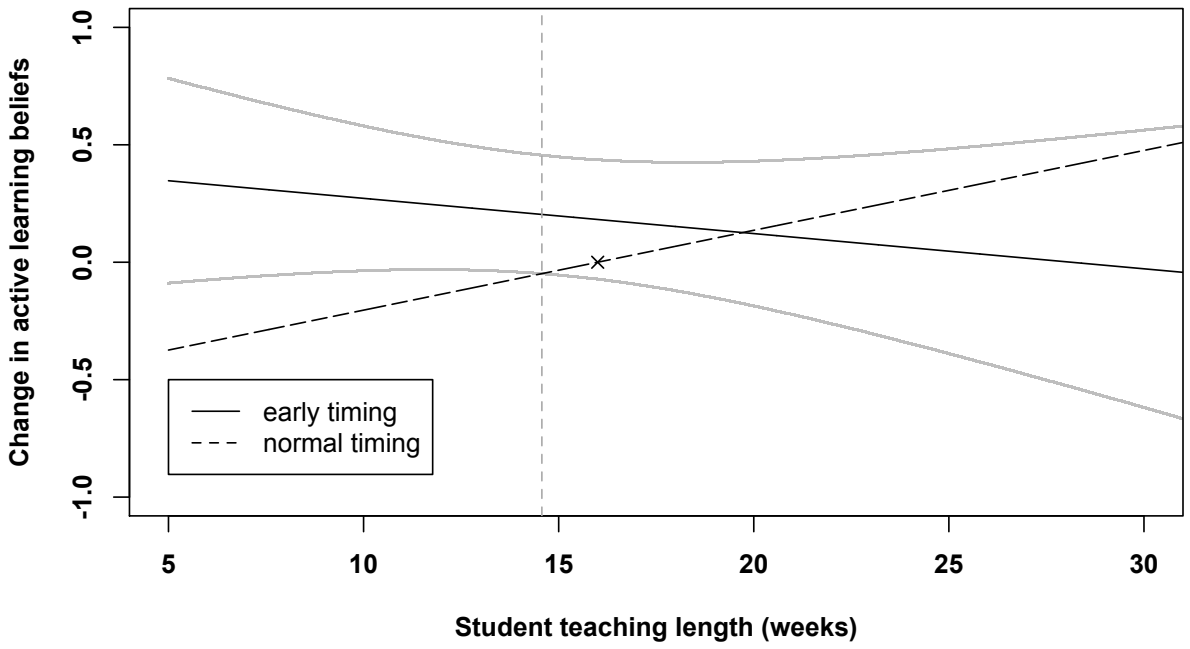


Figure 10. The effect of early timing of student teaching timing on active learning beliefs with 95% confidence bands as compared with predicted beliefs after 16 weeks of normally timed student teaching (marked ×); early timing had a statistically significant positive effect when student teaching was less than 14.6 weeks.

In summary, I found that program coherence was a significant positive predictor across all outcomes and that student teaching supervisor quality covariates were generally negative and not significant. The tertiary mathematics topics studied by the prospective teachers were statistically significant predictors in models of knowledge but not in models of beliefs. Turning next to the predictors of primary interest in the study—the quality, timing, and length of student teaching—I found that those variables were not significant predictors of pedagogical content knowledge. This result may have followed from the low program-level variance observed for that outcome, reflected by an ICC of .03 after including individual and program covariates. By contrast, the quality, timing, and length of student teaching were significant predictors of content knowledge and beliefs. For all three outcomes, student teaching length was found to moderate

the effect of timing, with early timing having significant positive effects only for prospective teachers in programs with shorter student teaching experiences. The quality of student teaching—measured as the frequency of practicum opportunities for connecting teaching and learning—was negatively associated with the prospective teachers' content knowledge and positively associated with both of their belief outcomes.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

In this chapter, I discuss findings from each study and from across the set of studies taken together. The findings have implications for policy concerning certification routes, the professional development of beginning teachers, and student teaching, and I address each of these topics in turn. The findings also have theoretical implications about the nature of teachers' knowledge and beliefs examined in these studies. In Chapter 1 and 2, I argued that those constructs (e.g., MKT, TSE) can be understood as components of mathematical proficiency for teaching, and so the results of the dissertation studies with respect to these constructs have implications for the nature of mathematical proficiency for teaching. These implications have consequences for how these and other components of mathematical proficiency for teaching are defined and measured, and I discuss measurement and how measures might be designed to better enable research on change in teachers' knowledge and beliefs. Finally, I discuss the implications of the findings for future research and especially for the design of an intervention positioned within student teaching.

Texas Study

One important finding of the Texas study was the statistical significance and large effect size of the perceived academic preparation measure across outcomes. This finding is somewhat surprising because it suggests that teachers' self-reports about their content-specific preparation in teacher education programs may have predictive utility for mathematical proficiency for teaching. For example, teacher education programs could use similar items at the end of the

program for evaluation. On the other hand, Hill (2010) argued that self-concept of mathematics (a similar self-report measure about mathematical ability) may not be useful as an outcome measure for professional development because it was only moderately correlated ($r = .25$) with mathematical knowledge for teaching (MKT). In this study, I found that academic preparedness to teach multiplicative reasoning topics was similarly correlated with MKT ($r = .27$) and slightly more highly correlated with the factors of teaching self-efficacy (TSE) ($r_{PE} = .30$, $r_{KE} = .47$). These results imply that although self-reports of academic preparation are strongly predictive of the outcome measures, they would not be sufficient to evaluate whether a teacher education program was producing teachers with mathematical proficiency for teaching. Certainly the outcome measures are only some of the components of the broad set that makes up mathematical proficiency for teaching, and moreover, the observed correlations are too low for self-reports of academic preparation to be reliable as replacements for these outcome measures.

Next I consider the results related to the student teaching variables. The first finding was that there was no effect of student teaching length or quality on the outcomes. Student teaching length ranged from 0 to 20 months, but that variation did not explain any of the variation in mathematical proficiency for teaching. The student teaching quality measure was also not significantly related to the outcomes. These findings are disturbing because student teaching is where teachers are supposed to integrate their study of theory and content with teaching practice. Ronfeldt and Reininger (2012) found that the quality (but not the length) of student teaching was positively related to teachers' reported feelings of preparedness to teach, teaching self-efficacy, and plans for teaching.

One limitation of the Texas study was the lack of adequate controls for selection bias. Limitations in the data set, including a small sample and potentially uncontrolled selection bias,

may be masking the true relationships. The finding that student teaching variables were nonsignificant predictors of mathematical proficiency for teaching demanded further exploration, particularly in a research design that would mitigate potential confounding variables. The U.S. study (Chapter 5) focused on student teaching and used a data set with numerous covariates including those similar to the set used by Ronfleldt and Reininger (2012).

I turn next to the teaching experience variables. I had hypothesized that the participants' collaborative activity related to student responses and reasoning would be related to their mathematical proficiency for teaching. The relationship was plausible because of pilot work and other research showing how teachers engaged in collaborative work can improve. For example, Koehler (2010) found that a school-based intervention called Instructional Consultation Teams positively affected the teaching self-efficacy beliefs of teachers in the intervention schools as compared with teachers in nonintervention schools in the same district. The intervention involved teachers working closely with other school personnel to solve an instructional problem with particular students. The measure of collaborative activity used in the Texas study would have picked up on the kind of work the IC Teams intervention might engender in a school—frequent collaboration with colleagues focused on student work and assessment data.

The predictive relationship in the Texas study between the variable representing collaborative activity and MKT or TSE lacked statistical significance, a surprising finding. The lack of significance may have been due to the nature of the measure used in the Texas study, which reflected the frequency of collaborative activities rather than their quality. Changing teachers' beliefs was an explicit goal of the IC Teams intervention (Koehler, 2010, p. 18), and teachers who work together habitually without external input on student assessment data, for example, may not have similar opportunities to experience change in their beliefs about their

own capabilities even though the activity might be similar in some respects to the activity entailed by the IC Teams intervention. The finding in the Texas study that collaborative activity did not have a significantly significant positive relationship with MKT or TSE suggests that frequency of collaborative activity is not the whole story when it comes to the role of teachers' collaboration in schools. In addition, the collaborative activity measure was not domain-specific. The elementary teachers might have collaborated frequently but might rarely have discussed mathematics teaching.

The length of teaching experience variable was significantly related to teachers' efficacy beliefs but not significantly related to their knowledge. This finding was surprising because Hill (2007, 2010) found that years of experience was a significant predictor of MKT. One difference is that the data set in the present study was restricted to teachers with 5 or fewer years of experience teaching, whereas both samples in the Hill studies were national samples that included veteran teachers. Perhaps the effect of experience is visible only over greater time spans.

Hill (2007) suggested that cohort bias could also explain the relationship between experience and MKT. Several decades ago, women had fewer professional opportunities, and more talented individuals might have become teachers in the past. As with Hill's study, the Texas study was not longitudinal, and therefore the teaching experience conflates comparisons across time with comparisons between cohorts. Participants in this study were hired between 2006 and 2010, a period that overlapped with the Great Recession. It is quite possible that teachers with different levels of experience (and thus hired in different years) were consequently dissimilar. The experience variable used in the Texas study accounts for the time people took the survey to mitigate that effect, but cohort bias could still drive the nonsignificance of results.

For example, even if the cohort hired in 2007 increased in mathematical proficiency for teaching over their first 2 years of experience, the mean outcomes for 2007 hires might be lower than the mean outcomes of those hired in 2009 if the weaker job market during the recession attracted more skilled applicants for the available teaching jobs.

Future research is needed to explore whether MKT changes over time and to confirm that domain-specific teaching self-efficacy develops with experience. In addition, to accurately investigate the relationship between experience and mathematical proficiency for teaching and to overcome the potential cohort bias in the present study, a longitudinal design is needed that would follow the same individuals as they gain experience. The development of mathematical proficiency for teaching over time was the central phenomenon addressed in the longitudinal Georgia study (Chapter 4).

Georgia Study

The main finding from the Georgia study was that mathematical proficiency for teaching did change over the semester of the study; on average, the teachers' MKT increased, and their personal teaching efficacy (PE) and knowledge efficacy (KE) decreased. The key challenge in interpreting the results of this study was to understand why PE and KE might have decreased over the semester even as MKT improved. One possibility is that the end-of-year exam functioned as a mastery experience, an opportunity for teachers to reevaluate their estimation of their own knowledge and ability to teach. Given some participants' comments about their experiences with the end-of-year exam in the previous year, and the apparent influence that exam had had on their self-efficacy beliefs, it is plausible that preparing for the test at the end of the semester depressed the self-efficacy scores in the last survey wave (early May).

Two other main groups of findings warrant discussion: the significant predictors of initial mathematical proficiency for teaching and the significant predictors of change. In contrast to the findings about MKT in the Texas study, teaching experience was statistically significantly and positively related to the teachers' initial status on all three outcomes including MKT. Having a secondary (Grades 6–12) teaching certificate was positively associated with the Georgia teachers' MKT and KE, but was statistically significant only in the model of MKT. This difference (and the finding of no relationship with PE) suggests that different factors of mathematical proficiency for teaching may stem from different kinds of preparation experiences. A secondary teaching certificate implies more mathematical training than an elementary certificate does, and it makes sense that teachers with that training might have higher MKT and KE scores. Their confidence in knowing how to teach multiplicative reasoning topics (PE) might not have been affected by their collegiate mathematical training.

Collaboration and experience teaching multiplicative reasoning topics were related to initial values in the Georgia teachers' MKT and PE, with interaction terms reaching (for PE) and nearly reaching (for MKT) statistical significance at the .05 level. The teachers with more frequent collaborative activity had a higher PE score for ratio and proportion topics than those who collaborated less frequently, but only if they were also teaching multiplicative reasoning topics. In contrast, the teachers involved in frequent collaborative activity had lower initial MKT scores than those who were not, unless they were also teaching ratio and proportion topics. This result might reflect school interventions in poorly performing schools (i.e., schools with less knowledgeable teachers might be more likely to have collegial activity focused on student thinking). Unfortunately, the estimated effect of school policies implementing this kind of intervention in these data (i.e., the estimate of collaboration on growth rate) was not significantly

different from zero. These findings suggest that collaborative activity has implications for mathematical proficiency for teaching, but that the relationships are complex. More research designed to untangle these relationships is needed before one can fully understand the nature of the relationship between collaborative activity and mathematical proficiency for teaching.

I next discuss the significant predictors of change in the outcome measures. The third model of MKT showed that early entry and mathematics professional development explained half of the observed variance in an individual's rate of change in MKT. The MKT growth over one semester teaching middle grades mathematics was significantly greater for early entry teachers, suggesting that MKT can be learned on the job by alternative route teachers. Somewhat surprisingly, more PD was negatively associated with the rate of change in MKT. I did not ask about the content of the PD, so these teachers may have been focused on learning in a different domain of mathematics that did not affect their MKT related to multiplicative reasoning topics.

Predictors in the third model of the PE outcome explained about one-fifth of the variation in PE rate of change observed across teachers in the Georgia study. As with the initial status for MKT and PE, collaboration activity interacted with the grade level at which the participants were teaching. The teachers who collaborated more frequently with others had a slower decrease in PE than those who collaborated less frequently, unless they taught in Grades 6 or 7. For Grades 6 and 7 teachers, more frequent collaboration was associated with a faster decrease in PE. Returning to the overall point about end-of-year test preparation as a mastery experience, I conjecture that collaborative activity among Grade 6 and 7 teachers during this same time functioned as social persuasion and a vicarious experience source of teaching self-efficacy. Discouraged teachers in these collaboration teams might have decreased the self-efficacy beliefs of other teachers.

The models of KE explored in the Georgia study did not explain the variation in the rate of change in KE over the semester. The only predictor that approached statistical significance (p = .78) was mathematics professional development, which had an ameliorative effect on the overall negative rate of change in KE. This estimate was surprising given the finding that professional development had a negative effect on MKT. Once again, the bigger message from these data is that these different outcomes for mathematical proficiency for teaching behave in different ways. Teacher education for prospective or practicing teachers may need to address each dimension of mathematical proficiency for teaching in different ways; there is little room left by these findings to hope for a panacea that would address all aspects of mathematical proficiency for teaching at once.

I turn now to the qualitative results of the Georgia study and the implications that arise when the results from the multilevel models of survey data and the qualitative interview data are integrated. Most obviously, the finding from the multilevel models that PE and KE decreased over the semester does not agree with the interviewed participants' accounts of steadily increasing confidence in their own mathematical knowledge of the curriculum and in their ability to teach effectively. One explanation for this discrepancy is that the end-of-year exam (and the teachers' work leading up to the exam reviewing content with their students) had a large effect on the measured PE and KE of the participating teachers. I have already described possible mechanisms (i.e., mastery experience or vicarious experience and social persuasion stemming from collaborative activity) that would account for the decrease. The teachers surveyed may have reported a generally increasing trend because they were looking at change over several years and perhaps even in reference to their first year of teaching.

This possibility suggests that there may be a cyclic, nonlinear relationship between teaching self-efficacy and time. For example, there might be an annual increase in teaching self-efficacy until the end-of-year exam, followed by a dip in self-efficacy beliefs once teachers received their students' scores. In this conjectured model, each year of experience would see an overall rise in MKT, PE, and KE together with a decrease in PE and KE during the second semester. Such a cyclic relationship would be consistent with the results from the interviews that described overall gains and with the surveys that described declines in PE and KE. One limitation of this longitudinal study is that it included only three points in time. To evaluate the possibility of nonlinear change, more than three points would be required, and to evaluate the possibility of overall year-to-year increases, the span of the study would need to be increased from one semester to two or more years.

United States Study

In the U.S. study, I aimed to estimate the effects of student teaching on the mathematical proficiency of prospective K–6 mathematics teachers. I hypothesized that high quality student teaching (i.e., providing opportunities to connect teaching with children's mathematical learning) that was timed early in the preparation program and thus concurrent with or followed by some of the content and methods courses would enable otherwise comparable prospective teachers in otherwise comparable programs to develop greater knowledge and more productive beliefs about teaching mathematics. The ability of the U.S. study to address this hypothesis was limited by the kind of measures of mathematical proficiency for teaching available in the TEDS-M data set (Tatto et al, 2012).

Observational survey-based studies such as the U.S. study one have many limitations for studying the possible effects of teacher education programs. The key limitation in this study was

an absence of what are frequently the best predictors of observed outcomes: pretreatment covariates for the outcome measures. Having such data and including them in the analysis would have increased the power of the study to detect significance in observed differences, and would have increased my confidence that selection bias had been adequately controlled. In addition, several important covariates (e.g., the program coherence and student teaching supervisor scales) and one substantive predictor (the quality of student teaching scale) were based on self-report data rather than more reliable data that might have been obtained through direct observation. Finally, the large amount of missing data in this study is not unusual for survey studies but remains problematic for analysis. I chose to impute missing data so that I could include almost all individuals in the sample. That choice reduced the ability of the analysis to detect significant results (i.e., standard errors in an analysis using listwise deletion would have been smaller), but that choice allowed more accuracy in the parameter estimates and produced results that could more validly be generalized to the population of Grades K–6 teachers being prepared in U.S. public institutions.

Another limitation of the U.S. study was that the student teaching—defined as "extended teaching practice" on the TEDS-M survey—was not directly observed. Direct discussion of children's thinking was one of the variables that was manipulated experimentally in Philipp et al.'s (2007) study, and field experiences that focused on children's mathematical thinking were more beneficial for teachers' knowledge and beliefs outcomes than field experiences that did not. The inferences in the present study were based on the assumption that the student teachers who were responsible for instruction would have many opportunities to attend to children's thinking. It is possible (even likely), however, that some student teachers responsible for instruction may not have noticed how their students were reasoning and responding during early student

teaching. The data set does not include data on what kinds of opportunities prospective teachers in early student teaching might have had to reflect on children's thinking and learning.

Despite these limitations, several well-warranted conclusions can be made from the findings of the U.S. study. First, the early timing of student teaching has a significant and positive relationship to teacher content knowledge and beliefs in programs with short student teaching (less than approximately 600 contact hours). These findings confirm (on a national scale) the results from Philipp et al.'s (2007) experiment in one teacher preparation program indicating that field experiences that are timed early can make a difference on teacher outcomes. Moreover, the results support the theoretical claim that experience of children's thinking is critical for teachers' learning.

The effect of timing on the prospective teachers' pedagogical content knowledge, however, was not significant, nor were the effects of student teaching length or quality. In fact, the only significant predictors of pedagogical content knowledge were tertiary mathematics coursework and prior secondary school achievement at the individual and program level. A related finding was that very little of the observed variability in pedagogical content knowledge could be attributed to differences in programs and none attributed to differences in student teaching. That finding suggests that all of the programs had similar effects on the prospective teachers' pedagogical content knowledge. Two options are possible: The programs did have a positive effect that was the same across all programs, or none of the programs had much if any effect. Because of the wide variety of programs in the study, I find it more plausible that the programs had little to no effect on the participants' pedagogical content knowledge than that they all had very similar effects. the available data, however, cannot support that claim. Preprogram scores on the pedagogical content knowledge measure would be necessary. Future research

should explore when and how pedagogical content knowledge develops in elementary teacher education programs.

Philipp et al. (2007) argued that their study provided evidence that prospective teachers would benefit from Dewey's laboratory approach in which teachers are guided in the careful exploration and study of children's mathematical thinking. These authors also recognized that establishing such an approach in teacher education programs might not be feasible because of the required cost and training for personnel. The U.S. study reported here provides some promising avenues for overcoming that obstacle: Timing student teaching before the last year and making it concurrent with or prior to content and methods courses might bring some of the benefit of the laboratory approach to prospective teachers. This timing would be particularly helpful if the early student teaching had an explicit focus on children's thinking—something that participants in the U.S. study may not have had. With programs that have constraints on the length of student teaching, simply scheduling some of the student teaching experience before the last year might be far more feasible that an intensive early field experience because only a minimal increase in resources would be needed. Future research—in particular randomized clinical trials—could be used to investigate the efficacy of such an intervention.

Looking Across the Studies

In this section, I discuss some implications of the findings from all three studies for teacher education. The relationship of early entry routes to certification and mathematical proficiency for teaching was explored in the Texas study and the Georgia study. I found no evidence in either study that early entry teachers have less MKT, PE, or KE than teachers who complete training before beginning to teach. In the Texas study, I did not find any differences in the relationships between other aspects of teachers' preparation (e.g., their perceived academic

preparation) and the outcome measures associated with early entry status. The best fitting model of mathematical proficiency for teaching did not distinguish among those groups of teachers. In the Georgia study, I found encouraging evidence that the MKT of early entry teachers changed more rapidly over the course of the semester than the MKT of other teachers—a finding suggesting that early entry status does not prevent and may even encourage learning on the job. The policy implications for the results from these two studies are clear: Early entry is a viable feature for teacher education programs, at least as far as the outcome measures used in these studies are concerned.

The results across the Texas and Georgia studies with respect to the professional development of beginning teachers are not as clear. I examined the possible effects of collaborative activity and professional development on mathematical proficiency for teaching. I found no evidence that collaboration makes a difference in the Texas study, and some evidence that collaboration has effects in the Georgia study. I also found mixed effects for mathematics professional development in the Georgia study: it decreased the rate of change in MKT and increased the rate of change in PE. The interview data from the Georgia study described a range of professional development experiences, but all experiences were short (just a day or two) and on the whole had limited perceived utility.

In contrast to the lackluster findings concerning professional development, there was strong qualitative evidence from interviews with some participants in the Georgia study that collaborating with mentors and other teachers (especially teachers with the same grade-level assignment) was of significant help to new teachers. These data warrant the recommendation to districts that new and novice teachers should be paired with mentors who teach the same subject and that collaborative activity should be encouraged. Such a recommendation must include the

caveat that such collaboration is only likely to be effective if the focus is improved teaching practice and student learning rather than a mechanism to distribute the labor of planning lessons or preparing activities.

Next, I turn to the implications of the Texas and U.S. study for policy concerning student teaching. I have already described the warrants from the U.S. study for recommending that teacher education programs with shorter (less than sixteen, 40-contact-hour weeks) schedule at least some of their student teaching earlier in the program. Such a change may increase prospective teachers content knowledge and beliefs. The lack of significant effects of student teaching on MKT or pedagogical content knowledge was disturbing. It certainly warrants further exploration, but also presents a challenge to teacher educators to find ways of designing student teaching to better support the transformative integration of content knowledge and pedagogical knowledge so that prospective teachers have better opportunities to learn content knowledge for teaching during their student teaching experience.

The results from all three studies have theoretical implications about the nature of mathematical proficiency for teaching and how it is formed and changes. All three studies have shown that the different factors of mathematical proficiency for teaching are not related to teachers' background or characteristics in the same way (TX and GA Study), and are not related in the same way to teachers experience on the job or in student teaching (GA and U.S. studies). In addition, the findings from the Georgia study suggest that teachers' domain-specific experiences (e.g., teaching ratio and proportion or teaching in Grades 6 or 7) can interact with their collaborative activity to affect initial status and rate of change with respect to domain-specific measures of mathematical proficiency for teaching. These findings mean that future work on mathematical proficiency for teaching should carefully specify how the constructs being

researched are defined and measured. Researchers should study teachers' characteristics with respect to different constructs (even subtly different constructs such as PE and KE) and how those characteristics might change as a function of different background and experiences. Mathematical proficiency for teaching may develop independently in different content domains of mathematics, such as multiplicative reasoning and geometry. Future teacher education efforts that aim to develop teachers' mathematical proficiency for teaching may require different experiences for the teachers for different outcomes. Different interventions and approaches tailored to each construct may be required.

Future Research

Each of the studies described in this dissertation invited several avenues for further research on mathematical proficiency for teaching. The findings from the Texas study led to the longitudinal Georgia study and the U.S. study of student teaching. The data collected for the Georgia study present an opportunity for another kind of analysis that would help further the conceptualization of mathematical proficiency for teaching. One important question is how the various components of mathematical proficiency might influence each other. Does knowledge influence teachers' knowledge efficacy, for example? Are there effects in the other direction? Cross-lagged structural equation models allow the investigation of these kinds of effects over time: Each outcome at time $n+1$ is modeled as a function of the other outcomes at time n. Differences in individuals can be controlled using covariates, and, for example, the effects of MKT on subsequent PE and KE could be estimated to provide evidence concerning the interdependence of knowledge and beliefs that make up mathematical proficiency for teaching.

The data used for the U.S. study offers similar opportunities for future research. I used the data from one country (the United States), and the international TEDS-M data set includes

data from 17 countries. A natural next step would be a comparative study of student teaching in other countries. In particular, the Philippines and Japan have similar variability in the timing of student teaching and have national mean outcomes below and above the United States, respectively. A comparative study of the three countries could ascertain the extent to which models that are viable in the U.S. sample would also describe the other two countries. Such a study would help explore the generalizability of the findings from the U.S. study.

A Word on Instruments

A major component of the Texas study was investigating the validity of domain-specific instruments of MKT, PE, and KE. The argument for the validity of these adapted instruments relied on the validity of the original instruments, which were designed for constructs that were broader or otherwise different. One limitation in the two studies that used these measures (Texas and Georgia) and of the U.S. study which used similar measures constructed using item response theory (IRT) comes from the assumptions imposed by the IRT model about how knowledge and beliefs might change.

The explicit goal of IRT is to create measures that are analogous to measures of physical quantities—just as a ruler is a measure of length. The standard error of measurement with IRT instruments can be quite large relative to observed variation in the scale (e.g. 0.3 to 0.5 SD on a scale with a range of 6 SD), so one should adjust the metaphor and further specify that the ruler has smudged markings. Or perhaps if the ruler is understood to be clearly marked, then one can only make measurements at dusk while looking through a dirty window pane.

Educational researchers (and others) might like to interpret increases in knowledge scores as indicative of learning, but what does knowledge change look like under the ruler metaphor? Under the metaphor, knowledge continuously increases from A units to become B units long and

passes through all the possible knowledge-lengths in between. It is not at all clear (indeed there is much evidence to the contrary) that conceptual change associated with learning happens in a way that is consistent with the ruler metaphor. Conceptual change is understood in scheme theory through jumps as children develop qualitatively different ways and means of operating to solve different classes of problems (e.g., Steffe & Olive, 2010). Conceptual change from a knowledge-in-pieces perspective involves the gradual refinement and increased coordination of a variety of knowledge resources that may all be "known" to the novice but in ways that are disconnected and inconsistent from the perspective of an expert (diSessa & Sherin, 1998; Wagner, 2006). Neither of these perspectives on conceptual change bears much resemblance to the naïve metaphor of accretion that IRT models foist upon those who use IRT instruments.

Fortunately, other psychometric options have recently been developed that hold more promise. One example is the diagnostic classification model (DCM), which reports the test-takers' "mastery" with respect to knowledge categories (e.g., Rupp, Templin, & Henson, 2010). The DTMR project used these models for the design of the DTMR instrument described in Chapter 2. DCMs may be more useful than IRT for large-scale research on conceptual change because DCMs do not make assumptions that the knowledge categories are ordered. In addition, mastery of a category implies a discrete jump in knowledge and does not specify how change happened, unlike the problematic implication of continuous accretion with IRT models.

I am concerned that studies that use IRT instruments to model change in knowledge may be making assumptions about the nature of conceptual change that are not warranted given current theory and research in mathematics education. Other kinds of assumptions are at the heart of the measures of teaching self-efficacy beliefs, and I consider those assumptions next.

The measures of teaching self-efficacy beliefs assume that the referent on an instrument has a similar meaning across participants. Suppose Teacher A agrees with the statement "I believe I am a good teacher," and Teacher B disagrees with the same statement. If "good teacher" does not have the same meaning for both teachers, their apparent disagreement may be illusory. When teaching self-efficacy is assessed for broad domains, such as teaching science, the risk of different meanings is lower and this issue is perhaps less problematic. When measures aim to focus on specific domains such as multiplicative reasoning, however, the question of how the items are interpreted becomes more important.

In the first two of the studies I reported, I was concerned with the domain of multiplicative reasoning. My definition of the domain (see Chapter 2) draws on perspectives that most teachers would not have access to, and teachers' conception of multiplicative reasoning may be quite different than my own. To get around this problem in the teaching self-efficacy measures, I used topics that can entail multiplicative reasoning (fractions, ratios, and proportions) rather than using the phrase "multiplicative reasoning." But this solution introduces a second problem. There are ways of conceiving of these topics that do not include reasoning of any kind; some teachers and some curricula used in the United States focus on procedures for solving problems but do not engage students in conceptual understanding of the problems or solutions.

Returning to the example of measuring teaching self-efficacy, suppose Teacher A agrees she is a good teacher of ratio topics (understanding ratio as a measure) and Teacher B disagrees that she is a good teacher of ratio topics (understanding ratio as a part-part comparison). In a second scenario, suppose Teacher C agrees that she is a good teacher (defining good teaching as procedural) before taking an extensive professional development course that changes her

understanding of what good teaching means. After professional development Teacher C disagrees that she is a good teacher, because she now interprets the question to be about teaching for understanding. In either scenario, comparing the scores tells us little about what is really different between Teacher A's and B's beliefs or about how Teacher C's beliefs about teaching changed.

The examples just described are hypothetical and somewhat extreme. The items I used in the Texas and Georgia studies included language that was less ambiguous about the tasks teaching involves; items asked about "answering student questions" and "monitoring student solutions." Yet even with more specific wording, interpretation of the language used in items on measures of beliefs may be problematic and warrants more careful investigation. To better study change in teachers' knowledge, instruments are needed that are well aligned to the hypothesized processes of change. Diagnostic classification models offer one promising solution because these models make fewer assumptions than IRT models about the nature of knowledge and how it changes. To better study teachers' beliefs, teachers' interpretation of items must be considered in validation work on new instruments and before the use of established measures, such as those used for measuring teaching self-efficacy beliefs in science or mathematics.

Designing an Intervention

I conclude this dissertation by sketching out the design of an intervention study that is informed by the findings I have discussed. The goal of the intervention is to increase prospective elementary teachers mathematical knowledge for teaching. I am particularly interested in MKT that belongs to the second and third categories I described in Chapter 2: knowledge to understand or appraise students' responses and reasoning and knowledge about the mathematical and instructional entailments of tasks and representations.

In the U.S. study I found that student teaching timed early in the program has positive effects on teachers' knowledge and beliefs. This finding motivated me to design an early student teaching experience that would supplement rather than replace traditional student teaching. The disturbing finding of a small to nonexistent effect of student teaching on pedagogical content knowledge motivated the goal of increasing MKT. By early student teaching, I mean the intervention would occur prior to the last year of a teacher education program. Such an intervention is different than an initial field experience because prospective teachers would plan and manage daily instruction over a monthlong period.

The intervention would take place in a semester-long content course on multiplicative reasoning. This course would draw on the tasks and sequencing used in current Number and Operation courses at the University of Georgia for middle and secondary teachers: an initial discussion of whole number multiplication and division, followed by fractions, and finishing with ratio and direct proportion. The first half of the course would be focused on that content.

The second half of the course would focus on children's mathematical thinking and would have two parts. The first part would be discussion of videos of children working on multiplicative reasoning tasks as in one of the laboratory approaches in Philipp et al.'s (2007) study of early field experiences. During this part each pair of prospective teachers would design 5 days of instruction focused on a specific topic from the content course (such as partitive division). All instruction would make use of the double number line representation. In the second part (a month long), prospective teachers would teach (and then revise) their lesson 4 times to 4 different groups of students. In light of the findings from the Georgia study concerning the value of collaboration, each pair of prospective teachers would meet daily and the class would meet

weekly to report on their instruction and collaborate on revisions in light of observed student reasoning and responses.

The monthlong student teaching portion of the course would provide a series of weeklong tutorial sessions for fourth- and fifth-grade children on a variety of multiplicative reasoning topics. Children participating in the tutorial would work with a pair of prospective teachers for a week before rotating to work with another pair on another topic. The instruction given by the prospective teachers would be constrained to have the same kinds of representations and language so that participating students would have a coherent experience. The tutorial could be scheduled during the school day or during an afterschool program, depending on the partner schools and other constraints. The instructor of the course and cooperating teachers would provide feedback to prospective teachers over the monthlong series of tutorials.

The intervention offers several opportunities for research. First, the intervention is promising as a means to increase prospective teachers' MKT because of the focus on children's thinking and because of the collaborative aspects; these hypotheses could be evaluated in a pilot study of the intervention. Second, the intervention can provide a site for research on how teachers' beliefs and knowledge change. Of particular interest is understanding the role of content, experience teaching, supervisor feedback, and collaboration in knowledge change. The monthlong tutorial section of the course may provide prospective teachers with mastery experiences that will help them develop positive beliefs about the content, their knowledge of the content, and their ability to teach the content. This hypothesis about how teachers' self-efficacy beliefs change could be examined during the intervention.

As I argued in the previous section, appropriate measures of teachers' beliefs and knowledge that are well aligned with the relevant theoretical explanations of how knowledge and

beliefs change are critical for research on teachers' knowledge and beliefs. Existing measures, including those developed and used in the three dissertation studies, leave many opportunities for improvement. Interviews conducted during the intervention would provide an opportunity to pilot items for use in large-scale measures. Should initial pilot work warrant scaling up the intervention beyond a single school of education, these measures would be needed for studies of effectiveness. Improved measures of teachers' knowledge and beliefs would have a wide range of applications in research on mathematics teacher education.

REFERENCES

Akar, G. K. (2010). Different levels of reasoning in within state ratio conception and the conceptualization of rate: A possible example. In P. Brosnan, D. B. Erchick, & L. Flevares (Eds.), *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 711–719).Columbus, OH: The Ohio State University.

Andrew, M. D. (1990). Differences between graduates of 4-year and 5-year teacher preparation programs. *Journal of Teacher Education, 41*(2), 45–51.

Andrew, M. D., & Schwab, R. L. (1995). Has reform in teacher education influenced teacher performance? An outcome assessment of graduates of an eleven-university consortium. *Action in Teacher Education, 18*(3), 43–53.

Armstrong, B., & Bezuk, N. (1995). Multiplication and division of fractions: The search for meaning. In J. T. Sowder & B. P. Schappelle (Eds.), *Providing a foundation for teaching mathematics in the middle grades* (pp. 85–119). Albany: State University of New York Press.

Asparouhov, T. & Muthén, B. (2010). *Chi-square statistics with multiple imputation*. Technical Report. Retrieved from www.statmodel.com/download/MI7.pdf

Baker, F. (2001). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Baker, F. & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Ball, D. L. (1990). Prospective elementary and secondary teachers' understanding of division. *Journal for Research in Mathematics Education*, *21*(2), 132–144.

Ball. D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 93*, 373–397.

Ball, D. L., Lubienski, S., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Ball, D. L., Thames, M., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education***,** *59*, 389–407.

Bandalos, D. L., & Finney, S. F. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). Hoboken, NJ: Routledge,

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice–Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Worth.

Barron's Educational Services. (2001). Profiles of American colleges (24th ed.). Hauppauge, NY: Author.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180.

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

Behr, M., Harel, G., Post, T., & Lesh, R. (1992). Rational number, ratio, and proportion. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 296–333). New York, NY: Macmillan.

Behr, M. J., Wachsmuth, I., Post, T. R, & Lesh, R. (1984). Order and equivalence of rational numbers: A clinical teaching experiment. *Journal for Research in Mathematics Education, 15*(5)*,* 323–341.

Ben-Chaim, D., Keret, Y., & Ilany, B. (2007). Designing and implementing authentic investigative proportional reasoning tasks: the impact on pre-service mathematics teachers' content and pedagogical knowledge and attitudes. *Journal of Mathematics Teacher Education, 10*, 333–340.

Berliner, D. C. (1994). Expertise: The wonder of exemplary performances. In J. Mangieri & C. C. Block (Eds.), *Creating powerful thinking in teachers and students: Diverse perspectives* (pp. 161–186). Fort Worth, TX: Harcourt Brace College.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*, 463–482.

Bill & Melinda Gates Foundation (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Seattle, WA: Author.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15.

Borko, H., Eisenhart, M., Brown, C., Underhill, R., Jones, D., & Agard, P. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education*, *23*(3), 194–222.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Boyle-Baise, M., & McIntyre, D. J. (2008). What kind of experience? Preparing teachers in PDS or community settings. In M. Cochran-Smith, S. Feiman-Nemser, & D. J. McIntyre (Eds.), *Handbook of research on teacher education* (3rd ed., pp. 307–330). New York, NY: Routledge.

Bradshaw, L., Izsák, A., Templin, J. & Jacobson, E. (under review). Diagnosing teachers' understandings of rational number: Building a multidimensional test within the diagnostic classification model framework. Submitted to *Educational Measurement: Issues and Practice*.

Brese, F., & Tatto, M.T. (Eds.). (2012). *TEDS-M 2008 user guide for the international database.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factorial covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Cabrera, A. F., & La Nasa, S. M. (2000). Understanding the college-choice process. *New Directions for Institutional Research*, *2000*(107), 5–22. doi:10.1002/ir.10701

Canada, D., Gilbert, M., & Adolphson, K. (2008). Conceptions and misconceptions of elementary preservice teachers in proportional reasoning. In O. Figueras, J. L. Cortina, S. Alatorre, T. Rojano, & A. Sepúlveda (Eds.). *Proceedings of the joint meeting of Psychology of Mathematics Education 32nd meeting and of the North American Chapter of the International Group for the Psychology of Mathematics Education 30th meeting* (Vol 2., pp. 249–256). Morelia, Mexico: Universidad Michoacana de San Nicolás de Hidalgo.

Carle, A. C. (2009, July). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*. doi: 10.1186/1471-2288-9-49.

Center for Research in Mathematics and Science Education. (2010). *Breaking the cycle: An international comparison of U.S. mathematics teacher preparation programs.* East Lansing: Michigan State University.

Chambers, S., & Hardy, J. (2005). Length of time in student teaching: Effects on classroom control orientation and self-efficacy beliefs. *Educational Research Quarterly*, *28*(3), 3–9.

Charalambous, C. (2009). *Preservice teachers' mathematical knowledge for teaching and their performance in selected teaching practices: Exploring a complex relationship* (Doctoral

dissertation). Retrieved from ProQuest Dissertations & Theses A&I. (Order No. 3343027)

Clift, R. T., & Brady, P. (2005). Research on methods courses and field experiences. In M. Cochran-Smith, & K. Zeichner (Eds.) *Studying teacher education: The report on the AERA panel on research and teacher education* (pp. 309–424). Mahwah, NJ: Erlbaum.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*, 778–820.

Cobb, V. (1999). *An international comparison of teacher education*. Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education.

Conderman, G., Morin, J., & Stephens, J. T. (2005). Special education student teaching practices. *Preventing school failure: Alternative education for children and youth*, *49*(3), 5–10.

Confrey, J. (1994). Splitting, similarity, and rate of change: A new approach to multiplication and exponential functions. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 291–330). Albany: State University of New York Press.

Confrey, J., & Smith, E. (1994). Exponential functions, rates of change, and the multiplicative unit. *Educational Studies in Mathematics, 26*(2), 135-164.

Confrey, J., & Smith, E. (1995). Splitting, covariation, and their role in the development of exponential functions. *Journal for Research in Mathematics Education, 26*(1), 66–86.

Cramer, K., Post, T., & Currier, S. (1993). Learning and teaching ratio and proportion: Research implications. In D. T. Owens (Ed.), *Research ideas for the classroom: Middle grades mathematics* (pp. 159–178). New York, NY: Macmillan.

Crocker L. & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education*, *51*, 166–173.

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

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education*, *53*, 286–302.

DiSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education, 20(10),* 1155–1191.

Drasgow, F., Levine, M. and Williams, E. (1985) Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. *Journal for Research in Mathematics Education*, *24*(1), 8–40.

Eisinga, R., Grotenhuis, M. & Pelzer, B. (2012). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 1–6, DOI 10.1007/s00038-012-0416-3

Empson, S. (1999). Equal sharing and shared meaning: The development of fraction concepts in a first- grade classroom. *Cognition and Instruction*, *17*(3), 283–342.

Empson, S., Junk, D., Dominguez, H., & Turner, E. (2005). Fractions as the coordination of multiplicatively related quantities: A cross-sectional study of children's thinking. *Educational Studies in Mathematics, 63*(1), 1–28.

Empson, S., & Turner, E. (2006). The emergence of multiplicative thinking in children's solutions to paper folding tasks. *Journal of Mathematical Behavior, 25*(1), 46–56.

Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics, 90*(8), 695–706.

Enochs, L., Smith, P., & Huinker, D. (2000) Establishing factorial validity of the mathematics teaching efficacy beliefs instrument. *School Science and Mathematics, 100*(4), 194–202.

Ensor, P. (2001). From preservice mathematics teacher education to beginning teaching: A study in recontextualizing. *Journal for Research in Mathematics Education*, *32*, 296–320.

Ericsson, K. A. (2004) Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70–S81.

Feiman-Nemser, S. (2001). From preparation to practice: Designing a continuum to strengthen and sustain teaching. *Teachers College Record*, *103*, 1013–1055.

Feistritzer, C. E., & Haar, C. K. (2008). *Alternate routes to teaching*. Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

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.

Fernandez, M. L., & Erbilgin, E. (2009). Examining the supervision of mathematics student teachers through analysis of conference communications. *Educational Studies in Mathematics, 72*, 93–110.

Fisher, L. C. (1988). Strategies used by secondary mathematics teachers to solve proportion problems. *Journal for Research in Mathematics Education*, *19*(2), 157–168.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., … Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*, 569–582.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. & Liddle, S. (2011). *The gateway to the profession: Assessing teacher preparation programs based on student achievement* (CEDR Working Paper 2011–2). Seattle: Center for Education Data & Research, University of Washington.

Goodson, I. F. (1993). Forms of knowledge and teacher education. *Journal of Education for Teaching*, *19*(4), 217–229.

Graeber, A., & Tirosh, D. (1988). Multiplication and division involving decimals: Preservice elementary teachers*'* performance and beliefs. *Journal of Mathematical Behavior*, *7*, 263–280.

Graeber, A., Tirosh, D., & Glover, R. (1989). Preservice teachers' misconceptions in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education*, *20*(1), 95–102.

Grossman, P. L., Valencia, S., Evans, K., Thompson, C., Martin, S., & Place, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literary Research, 32*, 631–662.

Hackenberg, A. J. (2007). Units coordination and the construction of improper fractions: A revision of the splitting hypothesis. *Journal of Mathematical Behavior, 26*(1)*,* 27–47.

Hackenberg, A. J. (2010). Students' reasoning with reversible multiplicative relationships. *Cognition and Instruction, 28*(4)*,* 383–432.

Hackenberg, A. J., & Tillema, E. S. (2009). Students' whole number multiplicative concepts: A critical constructive resource for fraction composition schemes. *Journal of Mathematical Behavior, 28,* 1–18.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in the public schools. *Journal of Economic Literature, 24*, 1141–1178.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research*, *66*, 397–409. doi:10.3102/00346543066003397

Harel, G., & Behr, M. (1995). Teachers' solutions for multiplicative problems. *Hiroshima Journal of Mathematics Education*, *3*, 31–51.

Harel, G., Behr, M., Lesh, R., & Post, T. (1994). Invariance of ratio: The case of children's anticipatory scheme for constancy of taste. *Journal for Research in Mathematics Education*, *25*(4), 324–345.

Harris, D. N., & Sass, T. (2007). *Teacher training, teacher quality and student achievement* (CALDER Working Paper 3). Washington, DC: Urban Institute.

Hart, K. (1981). Strategies and errors in secondary mathematics: The addition strategy in ratio. In C. Comiti & G. Vergnaud (Eds.) *Proceedings of the fifth conference of the International Group for the Psychology of Mathematics Education* (pp. 199–202).

Hart, K. (1988). Ratio and proportion. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 198–219). Reston, VA: National Council of Teachers of Mathematics.

Heck, D., Weiss, I., & Pasley, J. (2011). *A priority agenda for understanding the influence of the Common Core State Standards for Mathematics*. Chapel Hill, NC: Horizon. Retrieved from http://www.horizon-research.com/reports/2011/CCSSMresearchagenda/research_agenda.php

Heilig, J. V., & Jez, S. J. (2010). *Teach for America: A review of the evidence*. East Lansing, MI: Great Lakes Center for Education Research & Practice. Retrieved from http://greatlakescenter.org/docs/Policy_Briefs/Heilig_TeachForAmerica.pdf

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind policy initiative. *Educational Evaluation and Policy Analysis*, *29*, 95–114.

Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. *Journal for Research in Mathematics Education*, *41*, 513–542.

Hill, H. C. (2011). The nature and effects of middle school mathematics teacher learning experiences. *Teachers College Record*, *113*, 205–234.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794–831.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42,* 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004) Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11–30.

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–155). Charlotte, NC: Information Age.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Dirs.), *Test validity* (pp. 129–145*)*. Hillsdale, NJ: Erlbaum.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, *45*(7), 1–47.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *Elementary School Journal*, *93*, 355–372.

Humphrey, D. C., Wechsler, M. E., & Hough, H. J. (2008). Characteristics of effective alternative teacher certification programs. *Teachers College Record*, *110*, 1–63.

Izsák, A. (2008). Mathematical knowledge for teaching fraction multiplication. *Cognition and Instruction, 26*, 95–143

Izsák, A., & Jacobson, E. (2013, April). *Understanding teachers' inferences of proportionality between quantities that form a constant difference or constant product.* Paper presented at the National Council of Teachers of Mathematics Research Presession, Denver, CO.

Izsák, A., Jacobson, E., de Araujo, Z., & Orrill, C. H. (2012) Measuring mathematical knowledge for teaching fractions with drawn quantities. *Journal for Research in Mathematics Education*, *43*(4), 391–427.

Izsák, A., Tillema, E., & Tunç-Pekkan, Z. (2008). Teaching and learning fraction addition on number lines. *Journal for Research in Mathematics Education*, *39*(1), 33–62.

Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education, 41*(2), 169–202.

Jacobson, E. (2012). *Knowledge and personal efficacy for teaching and the sources of teaching efficacy for muliplicative reasoning*. Paper presented at poster session of the annual

meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Kalamazoo, MI.

Jacobson, E. (In press). Developing knowledge for teaching from experience: Mathematics teaching and professional development. In V. Ellis & J. Orchard (Eds.) *Learning teaching from experience: Multiple perspectives, international contexts*. London, UK: Bloomsbury.

Jacobson, E. & Izsák, A. (2012a). *Specialized content knowledge for teaching mathematical modeling in the middle grades*. Poster presented at the Studying the Emerging Challenges of the Common Core State Standards for Mathematics symposium, Columbia, MO.

Jacobson, E. & Izsák, A. (2012b). Using a knowledge-in-pieces approach to explore the illusion of proportionality in covariance situations. In Van Zoest, L. R., Lo, J.-J., & Kratky, J.L. (Eds.). *Proceedings of the 34th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, (pp. 629–636). Kalamazoo, MI: Western Michigan University.

Johnson, S. M., & Birkeland, S. E. (2008). Is fast-track preparation enough? It depends. In P. Grossman & S. Loeb (Eds.), *Alternative routes to teaching* (pp. 101–128). Cambridge, MA: Harvard Education Press.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based approach validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 129–164.

Kaput, J. J., & Maxwell-West, M. (1994). Missing-value proportional reasoning problems: Factors affecting informal reasoning patterns. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 235–287). Albany, NY: State University of New York Press.

Karplus, R., Pulos, S., & Stage, E. K. (1983). Early adolescents' proportional reasoning of "rate"problems. *Educational Studies in Mathematics*, *14*, 219–233.

Kennedy, M. (2008). Contributions of qualitative research to research on teacher qualifications. *Educational Evaluation and Policy Analysis, 30,* 334–367.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*, 568–589.

Kieren, T. E. (1988). Personal knowledge of rational numbers: Its intuitive and formal development. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 53–92). Reston, VA: National Council of Teachers of Mathematics.

Kieren, T. E. (1993). Rational and fractional numbers: From quotient fields to recursive understanding. In T. Carpenter, E. Fennema, & T. Romberg (Eds.), *Rational numbers: An integration of research* (pp. 49–84). Hillsdale, NJ: Erlbaum.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.) New York, NY: Guilford

Koehler, J. R. (2010). *An experimental evaluation of the effect of Instructional Consultation Teams on teacher efficacy: A multivariate, multilevel examination* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses A&I.

Kvale, S. (2007). *Doing interviews*. Thousand Oaks, CA: Sage.

Lamon, S. J. (1995). Ratio and proportion: Elementary didactical phenomenology. In J. Sowder & B. Schappelle (Eds.), *Providing a foundation for teaching mathematics in the middle grades* (pp. 167–198). Albany: State University of New York Press.

Lamon. S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework for research. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–668). Charlotte, NC: Information Age.

Levine, M. & Rubin, D. (1979) Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Lim, K. (2009). Burning the candle at just one end: Using nonproportional examples helps students determine when proportional strategies apply. *Mathematics Teaching in the Middle School*, *14*(8), 492–500.

Lobato, J., & Ellis, A. B. (2010). *Essential understandings: Ratios, proportions, and proportional reasoning.* Reston, VA: National Council of Teachers of Mathematics.

Lobato, J., Orrill, C., & Jacobson, E. (Under review). Investigating middle school teachers' proportional reasoning in the context of an assessment-development project. Submitted to *Journal of Mathematics Teacher Education*.

Lobato, J., & Siebert, D. (2002). Quantitative reasoning in a reconceived view of transfer. *Journal of Mathematical Behavior, 21*(1), 87–116.

Lobato, J. & Thanheiser, (2002). Developing understanding of ratio-as-measure as a foundation for slope. In B. Litwiller & G. Bright (Eds.), *Making sense of fractions, ratios, and proportions* (pp. 162 – 175). Reston, VA: National Council of Teachers of Mathematics.

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

Maandag, D. W., Deinum, J. F., Hofman, A. W. H., & Buitink, J. (2007). Teacher education in schools: An international comparison. *European Journal of Teacher Education*, *30*(2), 151–173.

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847–862.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Morris, D. (2010). *Sources of teaching self-efficacy: A scale validation* (Doctoral dissertation). Decatur, GA: Emory University. Retrieved from ProQuest Dissertations & Theses A&I. (Order No. 3423100)

Murray, S., Nuttall, J., & Mitchell, J. (2008). Research into initial teacher education in Australia: A survey of the literature 1995–2004. *Teaching and Teacher Education*, *24*(1), 225–239.

Musset, P. (2010). *Initial teacher education and continuing training policies in a comparative perspective: current practices in Organization for Economic Cooperation and Development (OCED) countries and a literature review on potential effects* (OECD Education Working Papers 48). Paris, France: Organization for Economic Cooperation and Development.

National Center for Education Statistics. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context*. Retrieved September 2, 2011, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Council for Accreditation of Teacher Education. (2010). *Transforming teacher education through clinical practice: A national strategy to prepare effective teachers* (Report of Blue Ribbon Panel on Clinical Preparation and Partnerships for Improved Student Learning). Washington, DC: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy.* Report by the Committee on the Study of Teacher Preparation Programs in the United States. Washington, DC: National Academies Press.

Newton, K. J. (2009). Instructional practices related to prospective elementary school teachers' motivation for fractions. *Journal of Mathematics Teacher Education*, *12*(2), 89–109.

Orrill, C. H., & Brown, R. E. (2012). Making sense of double number lines in professional development: Exploring teachers' understandings of proportional relationships. *Journal of Mathematics Teacher Education. 15*, 381–403. DOI 10.1007/s10857-012-9218-z

Pajares, F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 543–578.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: a comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*, 235–259.

Peressini, D., Borko, H., Romagnano, L., Knuth, E., & Willis, C. (2004). A conceptual framework for learning to teach secondary mathematics: A situative perspective. *Educational Studies in Mathematics*, *56*, 67–96.

Philipp, R. A. (2007). Mathematics teachers' beliefs and affect. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 257–315). Charlotte, NC: Information Age.

Philipp, R. A., Ambrose, R., Lamb, L. C., Sowder, J. T., Schappelle, B. P., Sowder, L., Thanheiser, E., & Chauvot, J. (2007). Effects of early field experiences on the mathematical content knowledge and beliefs of prospective elementary school teachers. *Journal for Research in Mathematics Education, 38*(5), 438–476.

Pitta-Pantazi, D., & Chritou, C. (2011). The structure of prospective kindergarten teachers' proportional reasoning. *Journal of Mathematics Teacher Education*, *14*(2), 149–169.

Post, T., Harel, G., Behr, M., & Lesh, R. (1991). Intermediate teachers' knowledge of rationale number concepts. In E. Fennema, T. Carpenter & S. Lamon (Eds.), *Integrating research on teaching and learning mathematics* (pp. 177–198). Albany: State University of New York Press.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Riley, K. (2010). Inservice teachers' understanding of proportional reasoning. In P. Brosnan, D. B. Erchick, & L. Flevares (Eds.). *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 4., pp. 1055–1061). Columbus, OH: The Ohio State University.

Rivkin, S. G., E. A. Hanushek, & J. F. Kain. (2005). Teachers, schools, and academic achievement. *Econometrica,* 73(2): 417–458.

Rizvi, N., & Lawson, M. (2007). Prospective teachers' knowledge: Concept of division. *International Education Journal*, *8*(2), 377–392.

Roberts, J.K., & Henson, R.K. (2000). *Self-efficacy teaching and knowledge instrument for science teachers (SETAKIST)*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education*, *28*, 1091–1106.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.

Saderholm, J., Ronau, R., & Brown, E. T. (2010). Validation of the Diagnostic Teacher Assessment of Mathematics and Science (DTAMS) instrument. *School Science and Mathematics, 110*, 180–191.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., . . . Schwille, J. (2007). *The preparation gap: Teacher preparation for middle school mathematics in six countries* (MT21 report). East Lansing, MI: Center for Research in Mathematics and Science Education. Retrieved from http:// usteds.msu.edu/MT21Report.pdf

Schwartz, J. L. (1988). Intensive quantity and referent transforming arithmetic operations. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 41–52). Reston, VA: National Council of Teachers of Mathematics.

Shechtman, N., Roschelle, J., Haertel, G., & Knudsen, J. (2010). Investigating links form teacher knowledge, to classroom practice, to student learning in the instructional system of the middle-school mathematics classroom. *Cognition and Instruction, 28*(3), 317–359.

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement, 5*(2/3), 70–80.

Seidman, I. (2006). *Interviewing as qualitative research (3rd ed.)*. New York, NY: Teachers College Press.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Silverman, J., & Thompson, P. (2008). Toward a framework for the development of mathematical knowledge for teaching. *Journal of Mathematics Teacher Education*, *11*, 499–511.

Silvernail, D. L., & Costello, M. H. (1983). The impact of student teaching and internship programs on preservice teachers. *Journal of Teacher Education*, *34*(4), 32–36.

Simon, M. (1993). Prospective elementary teachers' knowledge of division. *Journal for Research in Mathematics Education*, *24*(3), 233–254.

Simon, M. (2006). Key developmental understandings in mathematics: A direction for investigating and establishing learning goals. *Mathematical Thinking & Learning, 8*(4), 359–371.

Simon, M. A., & Blume, G. W. (1994). Building and understanding multiplicative relationships: A study of prospective elementary teachers. *Journal for Research in Mathematics Education, 25*, 472–494.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Smith, J. P., & Thompson, P. W. (2008). Quantitative reasoning and the development of algebraic reasoning. In J. J. Kaput, D. W. Carraher, & M. L. Blanton (Eds.), *Algebra in the early grades* (pp., 95–132). Mahwah, NJ: Erlbaum.

Son, J. (2010). Ratio and proportion: How prospective teachers respond to studnet errors in similar rectangles. In P. Brosnan, D. B. Erchick, & L. Flevares (Eds.). *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 4., pp. 243–251). Columbus, OH: The Ohio State University.

Sowder, J., Philipp, R., Armstrong, B., & Schappelle, B. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction: A research monograph*. Albany, NY: State University of New York Press.

Spooner, M., Flowers, C., Lambert, R., & Algozzine, B. (2008). Is more really better? Examining perceived benefits of an extended student teaching experience. *Clearing House, 81*(6), 263–270.

Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, *94*(2), 344–355.

Steffe, L. (1988). Children's construction of number sequences and multiplying schemes. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 119–140). Reston, VA: National Council of Teachers of Mathematics.

Steffe, L. (1993, April). Children's construction of iterative fraction schemes. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Seattle, WA.

Steffe, L. (2001). A new hypothesis concerning children's fractional knowledge. *Journal of Mathematical Behavior, 20*(3)*,* 267–307.

Steffe, L. (2003). Fractional commensurate, composition, and adding schemes learning trajectories of Jason and Laura: Grade 5. *Journal of Mathematical Behavior, 22*, 237–295.

Steffe, L. (2004). On the construction of learning trajectories of children: The case of commensurate fractions. *Mathematical Thinking and Learning,* 6, 129–162.

Steffe, L. & Olive, J. (2010). *Children's fractional knowledge*. New York, NY: Springer.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): Conceptual framework*. East Lansing: Teacher Education and Development International Study Center, College of Education, Michigan State University.

Tatto, M. T., Schwille, J., Senk, S.L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

Tchoshanov, M. (2010). Relationship between teacher knowledge of concepts and connections, teaching practice, and student achievement in middle grades mathematics. *Educational Studies in Mathematics*, *76*(2), 141–164.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127–146). New York, NY: Macmillan.

Thompson, P.W. (1994). The development of the concept of speed and its relationship to concepts of rate. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 181–234). Albany: State University of New York Press.

Thompson, P. W. (2011). Quantitative reasoning and mathematical modeling. In L. Hatfield, S. Chamberlain, & S. Belbase (Eds.), *New perspectives and directions for Collaborative Research in Mathematics Education,* (pp. 33–57). Laramie: University of Wyoming.

Thompson, P. W., & Thompson, A. G. (1994). Talking about rates conceptually, Part I: A teacher's struggle. *Journal for Research in Mathematics Education, 25*(3), 279–303.

Tourniaire, F., & Pulos, S. (1985). Proportional reasoning: A review of the literature. *Educational Studies in Mathematics, 16*(2), 181–204.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*, 783–805.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202–248.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*, 751–796.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, *27*(11), 1934–1943.

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 127–174). New York NY: Academic.

Vergnaud, G. (1988). Multiplicative structures. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in middle grades* (pp. 141–161). Reston, VA: National Council of Teachers of Mathematics.

Van Dooren, W., De Bock, D., Janssens, D., & Verschaffel L. (2008). The linear imperative: An inventory and conceptual analysis of students' overuse of linearity. *Journal for Research in Mathematics Education, 39*, 311–342.

Villegas-Reimers, E. (2003). *Teacher professional development: An international review of the literature*. Paris, France: International Institute for Educational Planning, UNESCO.

Wagner, J. (2006). Transfer in pieces. *Cognition and Instruction, 24(1),* 1–71.

Wang, A. H., Coleman, A. B., Coley, R. J., & Phelps, R. P. (2003). *Preparing teachers around the world*. Princeton, NJ, Educational Testing Service.

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

Wideen, M., Mayer-Smith, J., & Moon, B. (1998). A critical analysis of the research on learning to teach: Making the case for an ecological perspective on inquiry. *Review of Educational Research, 68*, 130–178.

Wilson, S. M., Ball, D., Bryk, A., Figlio, D., Grossman, P., Irvine, J. J., & Porter, A. (2009). *Teacher quality: Education policy white paper of the National Academy of Education.* Washington, DC: National Academy of Education.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education, 53*, 190–204.

Woolfolk, A. E., & Hoy, W. K. (1990). Prospective teachers' sense of efficacy and beliefs about control. *Journal of Educational Psychology*, *82*, 81–91.

Zeichner, K., & Gore, J. (1990). Teacher socialization. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 329–348). New York, NY: Macmillan.

APPENDIX A

CONCEPTUAL ANALYSIS OF MULTIPLICATIVE REASONING

In this section, I argue that a wide range of problems have the same mathematical and quantitative structure as whole number multiplication problems. I am not arguing that students or teachers do or should understand these problems in the way I will describe. I do believe, however, that this analysis shows the coherence of the domain of elementary multiplicative reasoning as defined in Chapter 2 and reveals promising opportunities for teaching and learning in this domain. Others have made similar arguments before (perhaps Vergaud, 1983 was first); my contribution here is to use Schwartz's (1988) categories of intensive and extensive quantities to distinguish fractions as numbers and fractions as ratios or rates and to in turn distinguish the multiplicative and quantitative structure of direct and inverse proportion problems. The extent to which the connections highlighted by this analysis are beneficial for mathematics teacher education is an open question under investigation at the University of Georgia.

The following analysis of multiplicative reasoning informs the instructional trajectory currently used in the Numbers and Operations content courses for middle grades and secondary teachers at the University of Georgia (Jacobson & Izsák, 2012a). In the analysis, I begin with whole number multiplication defined as grouping, then define fractions in terms of multiplication, use fractions and multiplication to describe two types of division, and finally demonstrate how proportions are a generalization of division. I also show how invariant product (inverse proportion) problems have mathematically and quantitatively distinct multiplicative structures.

The set of problem situations defining (elementary) multiplicative reasoning conform to what Schwartz (1988) called the (*I E E′*) semantic triad. Schwartz distinguished between extensive quantities (denoted *E* and *E′*), which can be directly counted or measured, and intensive quantities (denoted *I*), which cannot be counted or measured directly but are composed from other quantities. He claimed that "intensive quantity is essential to understanding the vast majority of situations that call for the arithmetic acts of multiplication and division" (p. 46).

The most common multiplicative structure is the (*I E E′*) semantic triad, which relates two extensive quantities and the intensive quantity defined by their quotient. For example, speed (*I*) is an intensive quantity defined as the quotient of time (*E*) and distance (*E′*); this definition is equivalent to the multiplication statement $I \times E = E'$ and gives rise to two division statements: $I = E'/E$ and $E = E'/I$. As I show below, these division statements are implicated in simple direct proportion problems.

I begin the content analysis by relating whole number multiplication understood as grouping (a common presentation in the early grades) to the (*I E E′*) semantic triad. Under this view of multiplication, the product (*P*) can be defined as the number of units in the collection of *G* groups such that there are *K* units in each group: $K \times G = P$. Consider the following multiplication problem: *Sam has 5 cans, and there are 3 tennis balls in each can. How many tennis balls does he have?* There are three quantities in this problem: the number of balls which is the product (*P*); the number of cans which is the number of groups (*G*); and the number of balls-per-can (*K*). The first two quantities are extensive and can be counted directly. The quantity of balls-per-can is intensive and is composed by taking the quotient of corresponding quantities of balls and cans, either from the given information (1 can contains 3 balls), from what Sam has in all (5 cans contain 15 balls), or from any hypothetical number of cans (including

partial cans) and balls (including partial balls). Therefore, K can be understood to represent an equivalence class [*K*], the rational numbers *pK/p*, where *p* is any whole number.

In Figure A1, I use a double number line to illustrate how whole number multiplication is an isomorphism of measures. The size of the product (denoted *P*) is measured in units, and the number of groups (*G*) that compose the product is measured in groups. These quantities can be understood as the two extensive quantities (denoted *E′* and *E,* respectively). The size of one group is given as *K* units, and the quantity *K* is measured in units. *K* is not the third (intensive) quantity in the semantic triad, but *K* can be used to construct the third quantity. The intensive quantity (denoted *I*) is represented with vertical dashed double-headed arrows. There is more than one such arrow because *I* is an equivalence class. In Thompson's (1994) language, each arrow is a *ratio* between a specific quantity of units and the corresponding quantity of groups. The equivalence class is a *rate*, a "reflected abstraction of constant ratio" (p. 7). The symbol *I* signifies all possible ratios between these measures as a single abstraction and thus highlights the invariant multiplicative comparison between units and groups. This multiplicative relationship defines the isomorphism between measures.

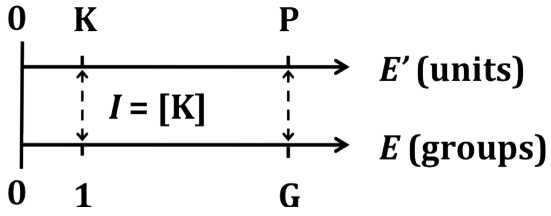


Figure A1. A double number line representation for the isomorphism of measures multiplicative structure.

The definition of multiplication can be used to define fractions, *A/B*. To define fractions for any denominator B, we interpret the unit as the product $B/B = 1/B \times B$, or *B* groups such that each group has the size of the unit fraction *1/B*. It follows that any fraction *A/B* can be

understood as the product of *A* groups such that each group has size *1/B: A/B = 1/B × A*. As before, there are three quantities but only two are extensive and can be measured directly. These quantities are (1) the product *A/B* which when measured is *A/B* times as long as 1 unit and (2) the number of unit fractions *A*, which is extensive because it can be counted.

The key for completing the analysis of fractions is understanding the unit fraction *1/B* as an intensive quantity. This understanding is difficult because the "units per group" status of this term in the product *A/B = 1/B × A* is easily confused with *1/B* as a number of units. We encountered the same difficulty in the multiplication example when interpreting *K* as the intensive quantity balls-per-can. The term *1/B* defines the size of each unit fraction and, as such, is a quotient between any quantity of *A/B* units and the corresponding number of unit fractions, *A*. This could, for example, be the quotient between 1 unit and the count of *B* unit fractions, or it could be the quotient between *1/B* units and the count of a single unit fraction. Therefore, *[1/B]* is an equivalence class, in the same way that *[K]* is an equivalence class (see Figure A2).

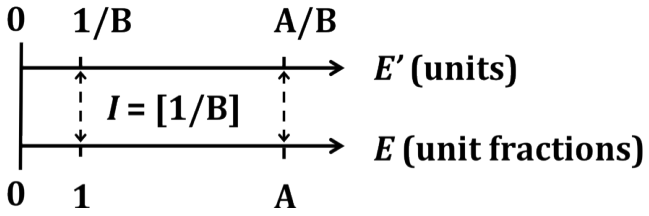


Figure A2. The multiplicative structure that defines the fraction *A/B* as an extensive quantity.

The definition of whole number multiplication by grouping can be extended to define fraction multiplication. To do that, we let *K*, *G = M/N*, and *P = Q/R* be fractions, and (as before) we define the product (*P*) as the number of units in *G* groups such that there are *K* units in each group: *K × G = P*. Note that the factors *K* and *G* have a different quantitative status; *K* is an intensive quantity (the equivalence class that defines an isomorphism between units and groups)

whereas *G* is an extensive quantity. Because *K* under whole number multiplication was already seen to be multiplicatively defined as an equivalence class, understanding *K* as a fraction in this more general definition of multiplication presents no difficulty. This notion of fraction as an equivalence class is closely related to the quantitative meaning of fraction as a ratio or rate, the multiplicative comparison between two quantities.

To make sense of the fractions *G* and *P* as extensive quantities, we use their respective denominators *N* and *R* to define two more isomorphisms of measures. The quantitative meaning for fraction as extensive quantities is closely related to the mathematical meaning of fractions as numbers—a single value defined in relation to 1. The first isomorphism relates the quantity of *groups* to the quantity of *unit fractions of groups* (size *1/N* groups) and is defined by the equivalence *[1/N]*, and the second isomorphism relates the quantity of *units* to the quantity of *unit fractions of units* (size *1/R* units) and is defined by *[1/R]* (see Figure A3). Each of the fractions *P* and *G* entail a multiplicative comparison between its magnitude and the magnitude of the constituent unit fractions, *1/R* or *1/N*.

The multiplication statement $K \times G = P$ yields two division statements that are distinct because of the role played by the intensive quantity *K*. The first type of division, $P/G = K$ is often called sharing or partitive division; it answers the quantitative question, How many are in each (one) group? There are two ways to understand the quotient. A child sharing 12 marbles among 3 friends equally might answer this question, "4 marbles" (see *X* in Figure A4a). Instead, we take the view that the answer is "4 marbles per friend" to stress that this quotient is an intensive quantity formed by composing any number of marbles and the corresponding number of friends. Thus under this view, the answer to a sharing division question is always an equivalence class (see *[X]* in Figure A4a).

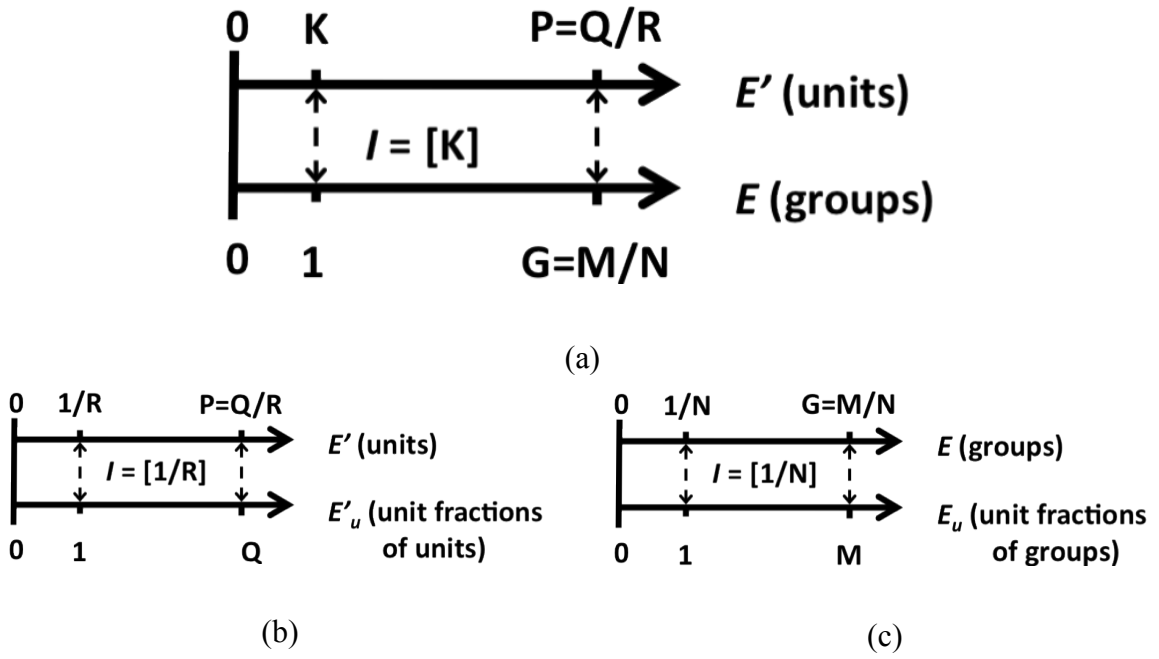


Figure A3. Multiplicative structures for (a) the product $P = Q/R$ of the fractions K and $G = M/N$, (b) the fraction $P = Q/R$, and (c) the fraction $G = M/N$.

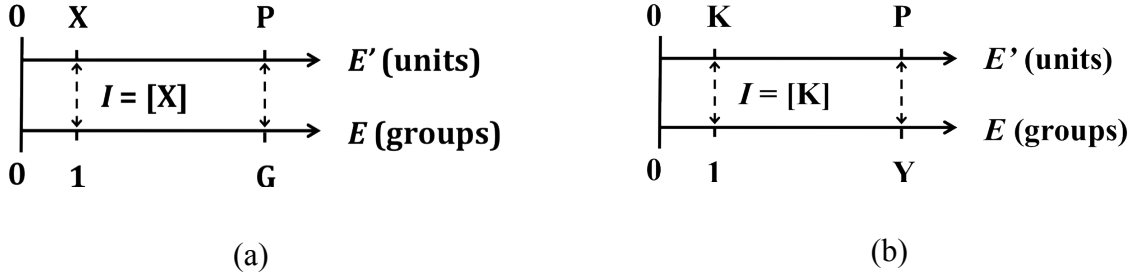


Figure A4. In (a) sharing division the quotient is an equivalence class *[X]*; in (b) measurement division the quotient is a quantity of groups *Y*.

The second type of division, $P/K = G$, is often called measurement or quotitive division. This form of division answers the quantitative question, How many groups are there? A measurement division problem in the marble context would be as follows: Twelve marbles are shared equally and each friend gets 4; how many friends got marbles? This kind of division

returns some number of groups as the quotient, and, like the previous kind can be illustrated with a double number line representation (see *Y* in Figure A4b).

Next I show how multiplicative comparisons can be used to relate simple proportions to the preceding discussion of division and multiplication. Let *E* and *E*' be two extensive quantities that covary proportionally, and let *A* and *C* be values of *E* that correspond to *B* and *D*, respectively, values of *E*'. We assume that three values are known and the fourth is unknown, and without loss of generality we can assume that either *C* or *D* is the so-called missing value. Partitive division yields $K = A/B$ (see Figure A5), the intensive quantity *I* that is a constant quotient between corresponding values of *E* and *E'*. There are many routes to a solution. The quantity *K* defines the homogeneous linear functions $f(E) = KE$ and $f'(E') = E'/K$ between the two extensive quantities. It follows that $A/B = C/D$ and that $B/A = D/C$. Moreover, several useful multiplicative comparisons (in this case, scale factors) can be made between values within each measure. For example, *D* is *D/B* times *B* and *C* is *C/A* times *A*. By the composition of multiplication, it follows that $C/A = D/B$. These relationships demonstrate that simple direct proportions share the same multiplicative structure as the arithmetic operations of multiplication and division and of fractions (and hence rational numbers).

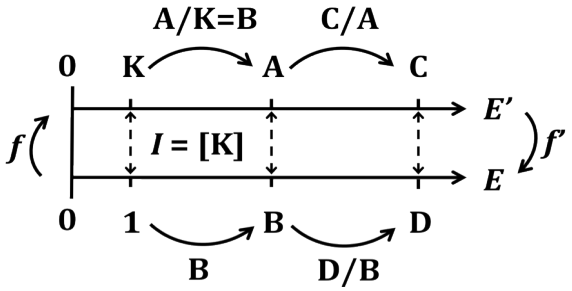


Figure A5. An elaborated illustration of the isomorphism of measures multiplicative structure for the simple proportion $A/C = B/D$ showing the rate or constant quotient *K*, the equality of the scale factors *C/A* and *D/B*, and the associated homogenous linear functions *f* and its inverse *f'*.

By contrast, invariant product situations involve a different multiplicative structure that Vergnaud (1983) called the product of measures. This structure "consists of the Cartesian composition of two measure-space … into a third" (p. 134). Schwartz (1988) categorized these problems with the ($E\ E'\ E''$) semantic triad, which had the defining multiplication statement $E \times E' = E''$. Schwartz observed that the referent for E'' is "entirely new" and "it remains to be defined along with its measure" (p. 51). There is some empirical evidence that this problem of quantification—defining the new extensive quantity and defining its measure as a Cartesian product—may be a significant challenge for preservice middle grades and preservice secondary teachers reasoning about these kinds of problems (Jacobson & Izsák, 2012b; Izsák & Jacobson, 2013). The multiplicative structure given by an isomorphism of measures or equivalently by the ($I\ E\ E'$) semantic triad is not adequate for reasoning about problems that involve an invariant product.

To make the point another way, consider the following example of an invariant product problem from Lamon (2007, p. 638): *If 3 people can mow and trim a lawn in 2 hours, how long will it take 2 people to do the same work?* The covarying quantities in this situation are all extensive: people, hours, and work, but for the sake of argument, I show how any corresponding pair of people and hours can be shoehorned into the multiplicative structure of the ($I\ E\ E'$) semantic triad. This multiplicative structure, however, cannot support the coordination between pairs that is the required to solve the problem.

Take the given pair of 3 people who finish the job in 2 hours. We begin by interpreting the number of people as an extensive quantity—after all, they can be counted. People will play the role of "groups" in this analysis. It is then necessary (but not very natural) to interpret *hours* as *person-hours per person*, an intensive quantity. It follows that the work accomplished is the

product 2 × 3 = 6 and has the appropriate unit of person-hour. This situation is illustrated in Figure A6a. The isomorphism between the measures of people and of work in the given case of 3 people and 2 hours is defined by the equivalence class *[K = 2]*. The same structure (isomorphism of measures) cannot accommodate the next pair of 2 people and 3 hours (the problem's answer), because for this next pair the isomorphism between work and people is defined by *[K = 3]* (see Figure A6b). The appropriate multiplicative structure for coordinating these two pairs can be represented with a Cartesian plane, as in Figure A6c). The invariant product can be seen in the areas of the rectangles defined by the origin and each person-hour coordinate pair. The locus of solutions is the graph of the function *f(people) = 6 / time*, which is equivalent to the quantitative equation $K \times G = P$.

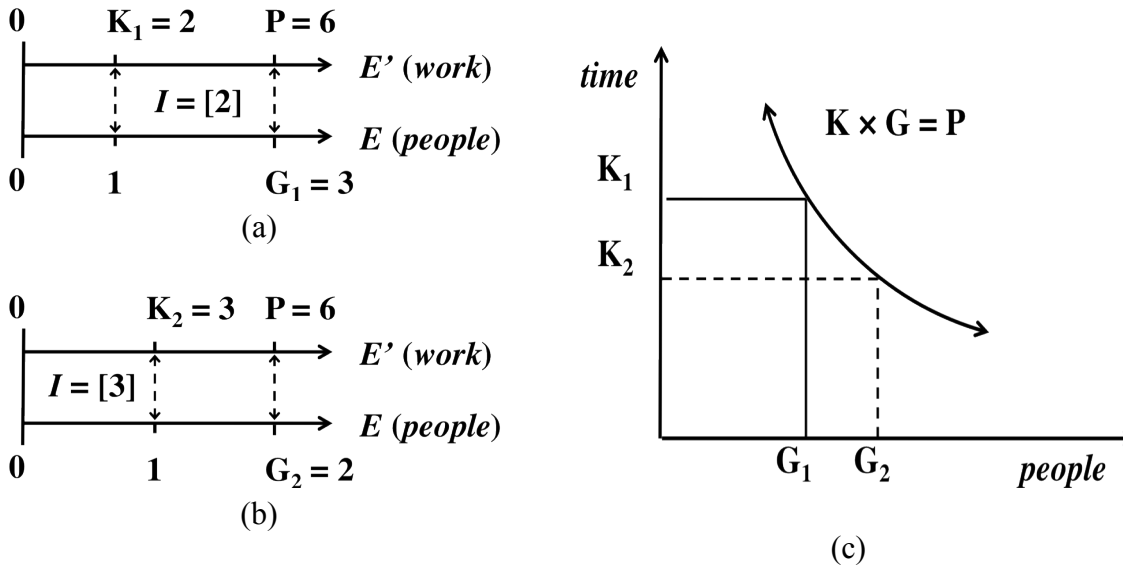


Figure A6. A multiplicative structure for (a) the 3 people, 2 hours pair; (b) a multiplicative structure for the 2 people, 3 hour pair; and (c) a multiplicative structure for $K \times G = P$ when all three are extensive quantities.

APPENDIX B

MATHEMATICAL KNOWLEDGE FOR TEACHING AND TEACHING SELF-EFFICACY INSTRUMENT DEVELOPMENT

In this appendix, I describe the details of the instrument development and revision process for the measures of mathematical proficiency for teaching multiplicative reasoning topics. First, I describe how I used item response theory (IRT) to develop and revise the MKT instrument and produce a short MKT instrument measuring the same construct. Next I describe the how I used confirmatory factor analysis (CFA) techniques to develop and revise the teaching self-efficacy beliefs instrument.

Mathematical Knowledge for Teaching Multiplicative Reasoning Topics

I used item response theory (IRT, e.g., Baker & Kim, 2004) to evaluate the strength of the MKT instrument and to obtain estimates of participants' latent trait ability (θ). Item response theory provides several advantages over classical test theory, including more robust methods for estimating θ in cases of missing data. I used the statistical package ltm available in R for the estimation of the IRT models discussed below. I also checked results by running equivalent models in MPLUS.

Missing Data

The response data used to calibrate the MKT instrument had a large amount of missing data owing to many not-reached items. The frequency of omissions is shown in Table A1. In summary, 83 individuals did not answer any of the MKT items, 58 individuals responded to only one item (25 missing values), and 238 individuals answered all of the MKT items. Missing data

on individual items ranged from 1.2% (Item 1) to 35.5% (Item 23); 18 items had less than 29% missing data. Missing data on item pairs ranged from 17.1% (Items 3 and 4) to 36.9% (Items 19 and 23). Missing data were handled in the analysis by using full information maximum likelihood (FIML) estimation, which is robust against bias and preferred over other missing data techniques (such as listwise deletion) because it uses all available data. Both the ltm package in R and MPLUS use FIML estimation by default.

Table B1

Numbers of Respondents With Omissions in the MKT Instrument Response Data

Number of missing responses						
0	1	2 to 8	9 to 16	17 to 24	25	All 26
238	22	15	27	49	58	83

Instrument Evaluation and Revision

First, I fit response data for 409 individuals on all 26 MKT items with a one-parameter logistic model (1PL) that estimates the difficulty (β) of each item but constrains all items to have the same discrimination parameter (α). In IRT, item difficulty is measured on the θ scale and is the ability level at which respondents have a 50 likelihood of answering the item correctly. The discrimination parameter quantifies how well an item distinguishes between high and low ability respondents. The 1PL model showed several indicators of misfit, including a significant parametric Bootstrap goodness-of-fit statistic using Pearson's χ^2 statistic ($p = .04$) in favor of rejecting the 1PL model and two items with significant χ^2 statistics ($\chi^2 = 18.00$, $p = .035$; $\chi^2 = 26.03$, $p = .002$) indicating item misfit.

Next, I fit response data on all 26 MKT items with a two-parameter logistic (2PL) model that estimated unique ability and discrimination parameters for each item. The 2PL model was

better fitting than the 1PL model; only one item exhibited significant model misfit ($\chi^2 = 20.54$, $p = .008$). The overall model goodness-of-fit test is not available for 2PL models in the ltm package. Instead, I compared the fit of the 2PL model with that of the 1PL model to see how much fit improved with the estimation of discrimination parameters. The information criterion indices (AIC and BIC) are based on the log-likelihood function and are used for model comparison. Models with smaller values are preferred. The BIC index penalizes the increase in parameters more severely than the AIC index; this difference was reflected in the contradictory indication provided by these indices about which model was better fitting (see Table A2). I also used a log-likelihood ratio test to compare model fit between the 1PL and 2PL models for MKT. The significant p value of the test ($p < .001$) provides evidence that the MKT items did not all share the same discrimination parameter, and so the 2PL model was statistically preferred.

Table B2

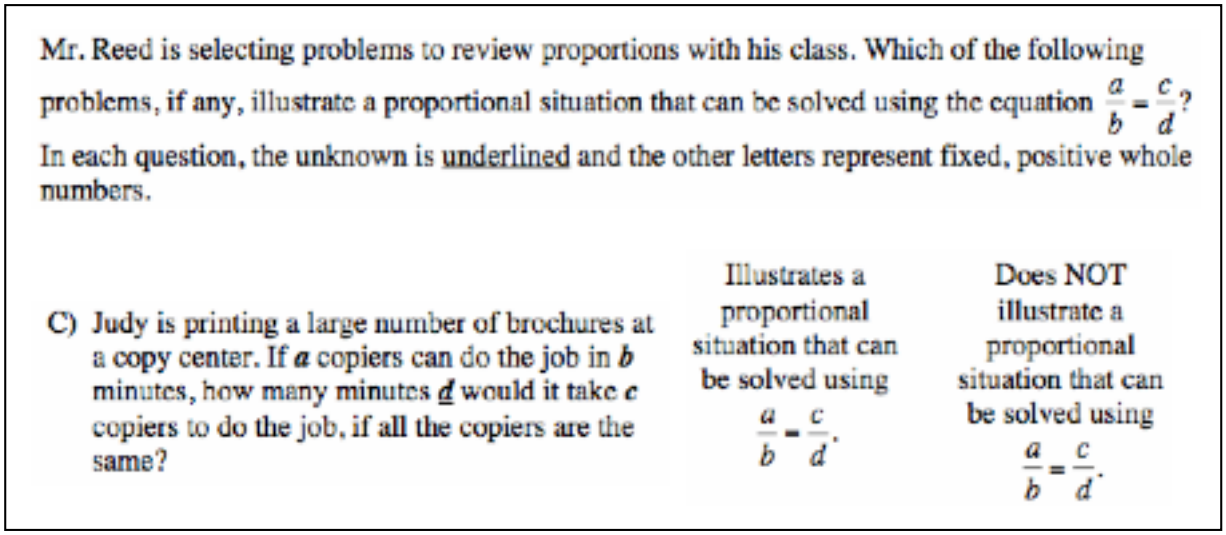
The Log-Likelihood Ratio Test and Model Fit Indices Comparing the 1PL and 2PL Models for All 26 MKT Items

Model	AIC	BIC	Log-likelihood	LRT	*df*
1PL	8113.81	8222.18	-4029.91		
2PL	8056.18	8264.89	-3976.09	107.63***	25

*** $p < .001$.

I was not satisfied with the 26-item 2PL instrument, because one item exhibited misfit. I examined the estimated parameters and found that Item 20 had an estimated difficulty of $\beta_{20} = 57.7$ *SD* and discrimination of $\alpha_{20} = .030$. I ran the same 2PL model in MPLUS and obtained similar item parameters ($\beta_{20} = 56.8$, $\alpha_{20} = .018$). The low discrimination can be interpreted to mean that Item 20 had very little ability to separate high and low ability respondents. The difficulty estimate suggests that it was extremely difficult. After I examined the item (see Figure

B1), I could see that it had been designed to tap an important aspect of teaching multiplicative reasoning—distinguishing quantities that maintain a constant quotient (direct proportion) from quantities that do not. It is plausible, however, that the content knowledge for teaching being tapped by Item 20 assessed a different dimension of MKT in the sense that responses on this item might vary independently of overall ability in the domain as measured by the other 25 MKT items. Prior research on with a similar problem (e.g., Fisher, 1988) has suggested that this kind of task is difficult for experienced teachers who might otherwise have strong content knowledge for teaching multiplicative reasoning topics. Answering this question correctly may require knowledge of a different multiplicative structure (the product of measures structure) than the one shared by all other items on the MKT instrument (isomorphism of measures, see Vergaud, 1983, 1988 and Appendix A). If that were true, then Item 20 would be unsuitable for a one-dimensional scale that included the other items.


Mr. Reed is selecting problems to review proportions with his class. Which of the following problems, if any, illustrate a proportional situation that can be solved using the equation $\frac{a}{b}=\frac{c}{d}$? In each question, the unknown is underlined and the other letters represent fixed, positive whole numbers.

	Illustrates a proportional situation that can be solved using $\frac{a}{b}=\frac{c}{d}$.	Does NOT illustrate a proportional situation that can be solved using $\frac{a}{b}=\frac{c}{d}$.
C) Judy is printing a large number of brochures at a copy center. If ***a*** copiers can do the job in ***b*** minutes, how many minutes ***d*** would it take ***c*** copiers to do the job, if all the copiers are the same?		

Figure B1. Item 20 on the MKT instrument, which did not function well with the rest of the items. (Copyright Erik Jacobson, 2013).

I revised the MKT instrument by removing Item 20 for two reasons. First, the item parameter estimates provided clear empirical evidence that Item 20 would not add precision to

the estimates of respondents' θ. Secondly, the risk that estimates of theta would be biased by removing Item 20 was mitigated because the test included several other items that dealt with the issue of appropriate multiplicative reasoning without introducing a novel multiplicative structure. For example, Item 18 asked respondents to distinguish quantities that maintain a constant difference versus a constant quotient (direct proportionality), and Items 23 and 26 required respondents to distinguish (invalid) additive and (valid) multiplicative reasoning strategies involving ratios.

Response data on the 25 MKT items remaining after removing Item 20 were fit to a 2PL model and evidenced no statistically significant item misfit at the .01 level. After evaluating item fit, I then examined person fit. The ltm package person-fit function computes the L_0 statistic of Levine and Rubin (1979) and the standardized analog L_z from Drasgow, Levine, and Williams (1985), including a *p* value for testing misfit. Of 409 respondents, 2 exhibited misfit at the .01 level, and 9 exhibited misfit at the .05 level. I removed those participants and re-estimated the 2PL model. The revised estimates of 398 respondents had only one person with misfit at the .05 level. Table B3 shows that there was minimal difference between the item parameter estimates obtained under each model and little difference between ability estimates either among the 11 individuals that originally exhibited misfit or across all respondents.

Final MKT Instrument.

The revised instrument (calibrated without the 11 misfits) was used as the final MKT instrument in the Texas study because person misfit can bias estimates of latent trait ability. The final instrument had high internal consistency (Cronbach's $\alpha = .95$). The final item parameters of the retained items are shown in Table B4. The item parameters were in an acceptable range (Baker, 2001). All items except one had difficulty parameters between -2.8 and 2.5 (Item 13 was

very hard, $\beta_{13} = 4.1$). Only one item had very low discrimination ($\alpha_1 = .24$), 5 items had low discrimination ($\alpha < .65$; Items 13, 16, 17, 18, & 26), and the remaining 19 items had moderate to very high discrimination. All items had point-biserial correlations greater than or equal to .2; with a sample size of approximately 400, these values are much higher than the recommended cutoff for inclusion of two standard deviations above 0 (Crocker & Algina, 2006).

Table B3

Increases in Item and Person Parameter Estimates From the Original to the Revised 25-Item 2PL Model

	Item parameter		Person ability (Δ_θ)	
	Discrimination (Δ_β)	Difficulty (Δ_α)	Misfits (n = 11)	All respondents (n = 409)
Maximum	0.334	0.129	0.096	0.123
3rd quartile	0.093	0.044	0.007	0.039
Median	0.018	0.005	-0.011	0.016
1st quartile	-0.005	-0.096	-0.028	-0.002
Minimum	-0.093	-0.802	-0.078	-0.091

Short MKT Instrument.

When I applied to do research in Georgia, some districts balked at the length of the surveys I proposed giving teachers. Negotiations with one large district (that in the end contributed 20% of the final sample) led to an agreement that I would shorten the MKT instrument and remove a number of background surveys and questionnaires. I analyzed the 25 successful items and selected 17 to cover the domain. Table B5 compares the content distribution of the MKT and short MKT instruments; the short MKT instrument retained a similar balance of content category and pedagogical tasks in the MKT instrument.

Table B4

Item Parameters for the Final MKT Instrument

Item	Difficulty (β)	Discrimination (α)	Point biserial correlation
1	2.457	0.24	.26
2	-1.887	1.58	.37
3	-2.646	0.68	.33
4	-1.760	2.27	.38
5	-1.294	1.01	.48
6	-2.655	1.35	.21
7	-0.188	0.68	.38
8	-1.147	0.95	.41
9	-0.518	1.99	.53
10	0.057	1.81	.57
11	0.145	0.77	.38
12	-0.644	0.67	.37
13	4.108	0.48	.26
14	-1.506	0.89	.35
15	-0.968	0.65	.36
16	-1.881	0.49	.32
17	1.322	0.62	.39
18	0.132	0.42	.33
19	-1.563	0.77	.36
21	-1.511	1.09	.40
22	-1.404	1.20	.44
23	-2.292	0.70	.34
24	-0.862	1.45	.44
25	-1.992	1.90	.31
26	-2.776	0.42	.30

I then analyzed a 2PL IRT model of these 17 items and found that the model fit well. Response data on the 17 MKT items evidenced no statistically significant item misfit at the .01 level. After evaluating item fit, I then examined person fit. I used the ltm package person-fit function to obtain *p* value for testing person misfit. Of 350 respondents (excluding some items meant that more people were missing data on all the rest), 2 exhibited misfit at the .01 level, and 9 exhibited misfit at the .05 level, a similar level of misfit as that calculated for the 25-item MKT model.

Table B5

Distribution of MKT Items by Pedagogical Task, Content Topic, and Problem Type on the MKT and Short MKT Instruments

	Pedagogical task		
Content topic	Understanding and appraising students' mathematical thinking	Selecting and using tasks and representations for instruction	Total
Proportional reasoning	MKT: 5 Short: 4	MKT: 2 Short: 2	MKT: 7 Short: 6
Fraction and ratio comparison	MKT: 9 Short: 5	MKT: 1 Short: 1	MKT: 10 Short: 6
Fraction multiplication and division	MKT: 4 Short: 3	MKT: 4 Short: 2	MKT: 8 Short: 5
Total	MKT: 18 Short: 12	MKT: 7 Short: 5	MKT: 25 Short: 17

The short MKT instrument had high internal consistency (Cronbach's $\alpha = .86$). The item parameters are displayed in Table B6. The item parameters were in an acceptable range (Baker, 2001). All items except two had difficulty parameters between -3.1 and 3 (Item 13 was very hard, $\beta_{13} = 4.8$; Item 3 was very easy, $\beta_3 = -3.6$). No items had very low discrimination; 5 items had low discrimination ($\alpha < .65$; Items 3, 13, 16, 17, 18), and the remaining 12 items had moderate to very high discrimination. All items had point biserial correlations greater than or equal to .2; with a sample size of approximately 350, these values are much higher than the recommended cutoff for inclusion of two standard deviations above 0 (Crocker & Algina, 2006). As a final check of the agreement between the MKT instrument and the short MKT instrument, I calculated the correlation between participants' MKT scores on these two instruments. It was extremely high, $r = .96$.

Table B6

Item Parameters for the Short MKT Instrument

Item	Difficulty (β)	Discrimination (α)	Point biserial correlation
3	-3.624	0.44	0.26
4	-1.736	2.30	0.40
5	-1.245	1.06	0.50
6	-3.049	1.00	0.22
7	-0.197	0.67	0.40
9	-0.534	1.92	0.56
10	0.043	1.68	0.59
11	0.128	0.83	0.45
13	4.781	0.39	0.27
16	-1.903	0.50	0.38
17	1.369	0.57	0.42
18	0.061	0.44	0.35
19	-1.537	0.76	0.39
21	-1.567	1.03	0.41
22	-1.412	1.13	0.45
24	-0.948	1.31	0.45
25	-2.229	1.36	0.30

Self-Efficacy for Teaching Multiplicative Reasoning Topics and Its Sources

As noted in Chapter 3, the Teaching Self-Efficacy Beliefs (TSE Beliefs) instrument was adapted from existing instruments to focus on practicing teachers (existing instruments focused on prospective teachers) and on teaching the content domain of multiplicative reasoning (existing instruments focused on teaching science). The general validity argument for the new instrument was based on the established validity and reliability of existing instruments (e.g., Hoy & Woolfolk, 1993). Researchers have been successful in modifying these items to focus on different school subjects (e.g., the Mathematics Teaching Efficacy Beliefs Instrument [MTEBI], Enochs, Smith, & Huinker, 2000, focused on mathematics rather than science) and on focused content (e.g., Newton, 2009, modified self-concept measures for mathematics to focus on fractions). Rather than using exploratory factor analysis to find a plausible factor structure of the

measures, I used MPLUS to estimate confirmatory factor analysis (CFA) models to determine whether the factor structure of the new instruments matched that of previous instruments. I also used CFA to determine whether the new instruments for the sources of teaching self-efficacy (which were adapted from existing instruments in the same manner) matched the four-factor structure predicted by theory and substantiated empirically with previously developed instruments.

The first important question was whether the TSE Beliefs instrument had a single-factor model (teaching self-efficacy for multiplicative reasoning) or a two-factor model (personal teaching efficacy and knowledge efficacy) in line with existing instruments (e.g., Roberts & Henson, 2000). The second question was whether the TSE Sources instrument evidenced a single-factor structure or the four-factor structure predicted by theory. Because these models are nested, I used a chi-square difference test to evaluate which model better fit the data and used overall indices of fit to evaluate whether the better model had adequate fit.

Missing Data

The response data for the TSE Beliefs instruments and the TSE Sources instruments included some missing data. As with the MKT instrument response data, the primary cause of missing data was not-reached items (i.e., participants stopped taking the survey before responding to all of the survey items). The pattern of missing data for the two instruments is presented in Table B7. For the TSE instruments, item omissions ranged from 0.2% (Item 3) to 1.1% (Item 15). On item pairs, omissions ranged from 0.4% (Items 3 & 7) to 1.8% (Items 8 & 12). For the Sources of TSE instruments, missing data on individual items ranged from .2% (Item A1) to 7.6% (Item S4). Missing data on item pairs ranged from .2% (Item A3 and Item A4) to 9% (Item S6 and M6). Because there were a relatively small number of missing values and

because the MLMV estimation method in MPLUS which corrects for nonnormal data required complete data (see below), I used listwise deletion to handle missing data when evaluating the TSE and TSE sources instruments.

Table B7

Number of Respondents With Omissions in Response Data for the TSE (Both TTMR Surveys, N=492) and TSE Sources Instruments (Second TTMR Survey, N=386)

	Number of missing responses				
Instrument	0	1	<50	<90	All
TSE Beliefs	426	15	2	2	47
TSE Sources	320	12	30	1	23

Assumptions

Most CFA estimation methods, including maximum likelihood estimation, assume that the data are multivariate normal. To assess that assumption, I computed the Shapiro-Wilk test for each TSE Beliefs and TSE Sources item; in all cases, the statistic was significant ($p < .001$) and led to rejecting the null hypothesis of univariate normality. This test is known to be extremely sensitive to small deviations from the normal distribution, so I next examined the skewness and kurtosis of each item. Kline (2005) suggested that absolute skewness should be less than 3.0 and absolute kurtosis less than 8.0 in structural equation models; these cutoff values were satisfied (see Tables A8 and A9). Next I examined Mardia's multivariate kurtosis coefficient using the mardia function from the R package psych. The multivariate kurtosis coefficient was 57.68 for the TSE Beliefs items and 50.17 for the TSE Sources items indicating multivariate non-normality.

Next, I examined at the Mahalanobis distance of individuals from the centroid. The square of the Mahalanobis distance is distributed as χ^2 with k degrees of freedom, where k is the number of variables. For both the TSE Beliefs data and the TSE Sources data, about 40 individuals had Mahalanobis distances beyond the χ^2 critical value at the .01 level of Type 1 error, suggesting that these individuals were multivariate outliers. Taken together, this information led me to conclude that that the TSE Beliefs and TSE Sources data were not multivariate normal. Instead of using the default MPLUS estimation method of maximum likelihood, I used the mean- and variance-corrected maximum-likelihood estimation (MLMV), which is designed to provide valid estimates when normality assumptions do not hold. Unfortunately, this method requires complete data, and therefore the subsequent analyses included only those individuals who answered all items ($N_{\text{TSE Beliefs}} = 426$ and $N_{\text{TSE Sources}} = 320$, $N_{\text{Both}} = 312$).

Instrument Evaluation and Revision

In a preliminary analysis involving the first 266 TTMR survey participants, I identified a subset of items for each source that (a) had high (> .5) standardized factor loadings, (b) had moderate to high (> .3) r^2 values indicating the variance in each item explained by the latent trait, and (c) maintained the highest possible Cochran's alpha coefficient of reliability (Jacobson, 2012). Removing items that did not fit these criteria led to more parsimonious models for each factor of the TSE Sources instrument without sacrificing the overall psychometric quality. Descriptive statistics for these sets of items are presented in Tables B8 and B9.

One factor of TSE Sources—vicarious experience—had an unacceptable alpha coefficient of .58 in the preliminary study. Researchers have had difficulty constructing internally consistent instruments for measuring vicarious experience as a source of other self-

efficacy constructs and have noted that vicarious experiences from peers and those from self may function as separate subfactors (Usher & Pajares, 2009). After I examined a preliminary model, I saw that Items v3 and v4 were the cause of the model misfit, and an analysis of the wording of the items suggested that Items v3 and v4 were about vicarious experience from peers (e.g., Item v4: "By watching excellent teachers around me, I often learn better ways to approach my own teaching of topics involving fractions, ratios, and proportions."), whereas Items v5 and v6 had to do with vicarious experiences from self (e.g., Item v5: "When I am preparing to teach topics involving fractions, ratios, and proportions, I often try to visualize myself working through the most difficult teaching situations."). Moreover, Items v3 and v4 did not correlate strongly with each other ($r = .44$) or with the other two items ($r < .24$), but Items v5 and v6 had a significant correlation at $r = .648$. The Spearman-Brown reliability is more appropriate than Cronbach's alpha for two-item scales (Eisinga, Grotenhuis, & Pelzer, 2012), and Items v5 and v6 alone had a Spearman-Brown reliability coefficient of .79. To minimize measurement error and to meet minimum reliability criteria for a measurement factors, I used a simplified instrument for the sources of vicarious experience that included only Items v5 and v6.

Using additional response data from the full analytic sample, I was able to confirm the expected factor structure for the TSE Beliefs instrument and for the TSE Sources instrument. In addition, I was able to identify two misfitting items in the vicarious experience factor and find a substantive explanation for the misfit. Removing these items increased the coefficient alpha for the vicarious experience factor to .78. The final CFA models show that both the TSE Beliefs and the TSE Sources instruments are well behaved psychometrically.

Table B8

Univariate Statistics for Item Data From the TSE Beliefs Instrument

TSE Beliefs item	*M*	*SD*	Skewness	Kurtosis
Personal efficacy				
2	3.84	0.97	-0.68	0.12
5	4.05	0.80	-1.10	1.24
7	3.85	0.82	-0.82	0.10
8	4.00	0.95	-0.92	-0.14
11	4.28	0.84	-1.23	1.59
Knowledge efficacy				
3	3.79	1.08	-0.94	1.61
4	3.92	1.18	-0.52	0.26
6	4.01	1.00	-1.20	1.46
9	4.25	0.77	-1.11	0.91
10	4.05	0.92	-1.07	1.86
12	4.38	0.76	-1.26	1.83
15	4.23	0.79	-0.90	1.00

To evaluate the TSE Beliefs instrument, I fit a single factor CFA model and compared it with the theoretically predicted two-factor CFA fit using the same data. Because these models were nested, I used a log-likelihood ratio test ($\chi^2 = 144.08$, $df = 1$, $p = .000$) and concluded that I could reject the null hypothesis of no increase in model fit due to the extra factor. To evaluate the TSE Sources instrument, I fit a single factor CFA model and compared it with the theoretically predicted four-factor CFA fit using the same data. Again, I used a log-likelihood ratio test ($\chi^2 = 651.33$, $df = 6$, $p = .000$) and concluded that I could reject the null hypothesis of no increase in model fit due to the extra factors. These results provide a strong confirmation that these measures have a factor structure analogous to that of the measures from which they were adapted and therefore are well aligned with Bandura's (1997) social cognitive theory.

Table B9

Univariate Statistics for Item Data from the TSE Sources Instrument

TSE Sources item	*M*	*SD*	Skewness	Kurtosis
Mastery experience (ME)				
m2	4.22	0.76	-1.09	1.83
m3	3.86	0.99	-0.80	-0.02
m4	3.83	0.94	-0.75	0.27
m5	4.19	0.81	-1.23	2.22
Social persuasion (SP)				
s2	4.19	0.65	-0.48	1.19
s3	2.95	1.02	0.20	-0.15
s4	3.81	0.75	0.00	-0.36
s5	3.76	0.87	-0.65	0.84
s6	3.65	0.82	-0.23	0.09
Vicarious experience (VE)				
v3[a]	4.09	0.84	-0.85	0.87
v4[a]	3.45	0.95	-0.76	0.38
v5	3.52	1.01	-0.49	-0.41
v6	3.74	0.88	-0.63	0.3
Emotional & physiological states (EP)				
e3	3.83	1.12	-0.79	-0.27
e4	4.07	0.92	-1.09	0.97
e5	4.04	1.02	-1.08	0.51
e7	4.04	1.06	-1.19	0.84

[a] These items were removed from the final scale because of item misfit.

Finally, to further evaluate the validity of the instruments, I fit a structural equation model in which both factors of TSE were regressed on the four sources of TSE. To check the fit of the models, I used the Satorra–Bentler (SB) χ^2 test statistic, which is appropriate when data are nonnormally distributed (Kline, 2005). The two-factor TSE models, the four-factor TSE sources model, and the structural equation model of sources predicting TSE all exhibited significant χ^2, which led to rejection of the null hypothesis that the model fits the data (Table B10). Scholars widely agree (e.g., Kline, 2005; Bandalos & Finney, 2010), that the χ^2 is overly sensitive and in practice rely on fit indices instead. I used three other indices to determine the fit of each CFA and structural equation model: the comparative fit index (CFI); the root mean

square error of approximation (RMSEA); and the standardized root mean square residual (SRMR). Kline (2005) stated that CFI less than .95, RMSEA greater than .05 with 95 confidence interval exceeding the interval of 0 to .10, and SRMR greater than .08 can all indicate model misfit. Table B10 lists the fit indices for each model. All fit indices for each of the preferred models were within the acceptable range, except the CFI index for the model of sources predicting TSE, which at .94 was very close to the accepted cutoff. AIC and BIC are relative fit criteria (described previously), and smaller values indicated that the 2-factor and 4-factor models were preferred over their 1-factor counterparts.

Table B10

Model Fit Indices for the CFA Models of the Teaching Self-Efficacy (TSE) and TSE Sources Instruments and for the SEM Regressing TSE on Its Sources

Model	χ^2_{SB}	*df*	CFI	RMSEA [a]	SRMR	AIC	BIC
TSE Beliefs							
(1 factor)	190.5 $p = .000$	54	.91	.077 (.06, .09)	.054	10863	11009
(2 factors)	116.4 $p = .000$	53	.96	.053 (.04, .07)	.037	10720	10871
TSE Sources							
(1 factor)	409.1 $p = .000$	90	.70	.105 (.09, .12)	.113	11012	11182
(4 factors)	110.1 $p = .030$	84	.98	.031 (.01, .05)	.047	10373	10565
TSE Sources and TSE Beliefs	409.7 $p = .000$	309	.94	.032 (.02, .04)	.050	17192	17551

[a] The 90% confidence interval for the RMSEA fit index is indicated within parentheses.

The final measurement models for the TSE Beliefs instrument and for the TSE Sources instrument are presented in Figures B2 and B3. Each factor of TSE Beliefs and TSE Sources was

modeled as a latent (unobserved) variable. Recall from Chapter 3 that latent variables are represented in SEM diagrams with an oval. Each item was modeled as a manifest (observed) indicator and is represented by a square. The CFA measurement framework postulates a unique latent error term for each item. For example, e2 in Figure B2 is the latent error term for the PE item t2. The path coefficients from these latent error terms to the items can be interpreted as estimates of measurement error, the proportions of variance in the observed item scores that is not explained by the latent factor the item loads onto. Arrows indicate the predictive paths between latent and manifest variables. Path coefficients can be read as regression coefficients; for every unit increase in the variable at the tail of a path, we can expect an increase in the variable at the head of the path of magnitude equal to the number of units represented by the path coefficient. Standard errors are given in parentheses in the diagram in Figures B2 and B3, and *p* values can be calculated for each coefficient by dividing the coefficient by the standard error to obtain the appropriate *z*-score

All standardized loadings for both models were significant at the $p = .01$ level and ranged in magnitude from .65 to .85 on the TSE Beliefs instrument and from .62 to .89 on the TSE Sources instrument. In the TSE Beliefs measurement model, personal efficacy and knowledge efficacy were significantly correlated at $r = .80$. The four factors of the TSE Sources instrument were allowed to correlate, and all correlation estimates were significant and ranged from $r = .15$ between vicarious experience and emotional and physiological states to $r = .75$ between mastery experiences and emotional and physiological states. In the TSE Beliefs measurement model, the minimum r^2 was .33, and the r^2 for half of the items (6 items) was greater than .56. In the TSE Sources measurement model, the minimum r^2 was .38, and the r^2 for half of the items (8 items)

was greater than .56. These results confirm that both the TSE Beliefs and the TSE Sources instruments were functioning as expected under social cognitive theory.

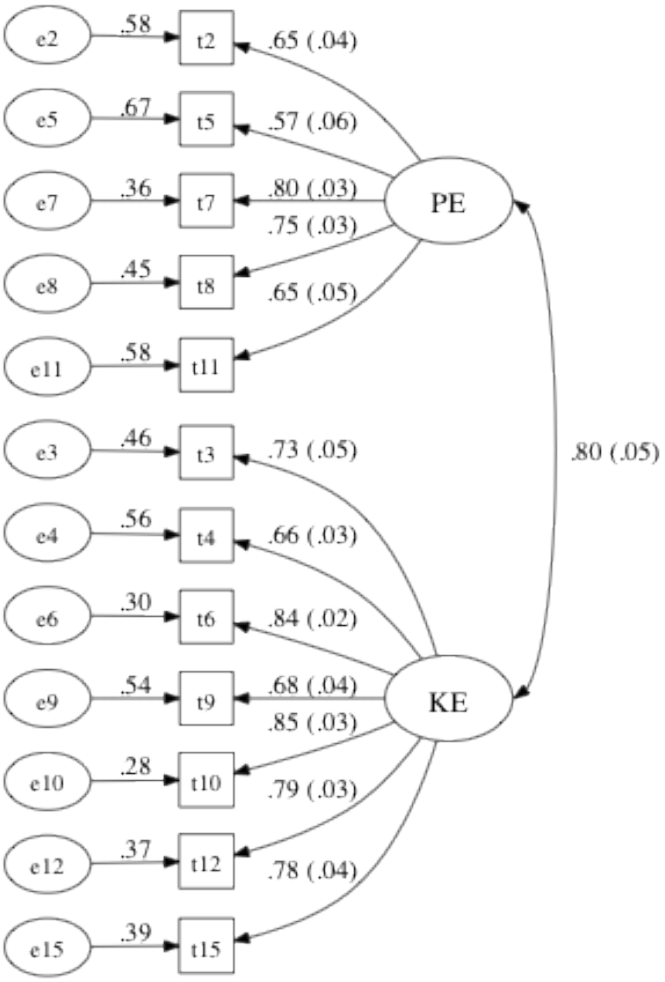


Figure B2. Measurement model for the 12-item Teaching Self-Efficacy instrument. All path coefficients were statistically significant at $p < .01$.

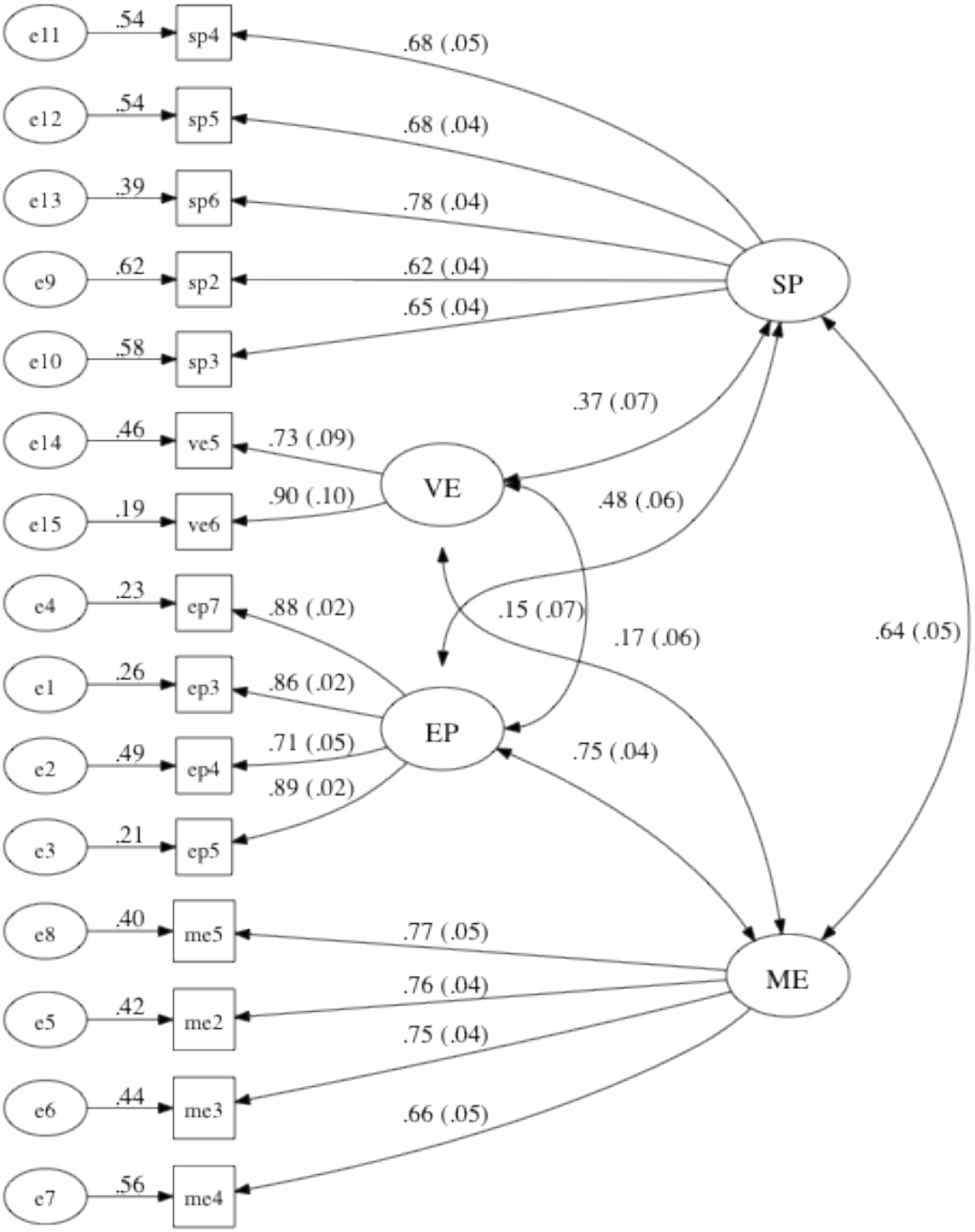


Figure B3. Measurement model for the 15-item Sources of Teaching Self-Efficacy instrument. All path coefficients were statistically significant at $p < .01$.

APPENDIX C

INTERVIEW PROTOCOL FOR THE GEORGIA STUDY

1. Tell me about how you came to be a mathematics teacher: Why teaching? Why math?
2. I'm interested to learn more about how you teach.
    a. What are your mathematical goals for students? Have these changed since you started teaching? How?
    b. What do you do to support the students that are doing well? What about the students who are not doing well? Have these strategies changed since you started teaching? How?
    c. Do you use manipulatives? Do you use partial credit? How and when? Has your use changed since you started teaching? How?
3. Can you think of a time that a student's mathematical work surprised you? Why was it surprising? How did you handle the situation? If you encountered the situation again, would you do the same thing? Would you have done the same when you were a beginning teacher?
    a. Can you think of a time when you knew a student had a misconception. How did you know? What did you do in response? If you encountered the situation again, would you do the same thing? Would you have done the same when you were a beginning teacher?
4. How confident are you that you know the mathematics well enough to teach it? How confident are you that you know how to teach math well? Has your confidence changed since you started teaching? Why?
5. Do you remember any activities with other teachers that helped you learn how to interpret student responses? What were they and how did they help?

a. Do you remember any activities with other teachers that helped you improve as a teacher? What were they and how did they help?

b. Did you ever work with a mentor teacher or coach? In a grade level planning team or subject-area department?

6. Can you remember any formal learning experiences—such as college classes or professional development workshops—that helped you learn how to interpret student responses? Do you remember any that helped you improve as a teacher? What were they and how did they help?

7. Last question: Of all the things we've talked about—and anything you might have remembered but that we didn't get a chance to talk about— what past experiences have made the biggest difference in becoming a better teacher?