5-1-2010

Mathematics self-efficacy and calibration of students in a secondary mathematics teacher preparation program

Joseph Champion
University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

THE MATHEMATICS SELF-EFFICACY AND CALIBRATION OF STUDENTS IN A SECONDARY MATHEMATICS TEACHER PREPARATION PROGRAM

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Joseph Keith Champion

Natural and Health Sciences
School of Mathematical Sciences
Educational Mathematics

May, 2010

This Dissertation by: Joseph Keith Champion

Entitled: *The Mathematics Self-Efficacy and Calibration of Students in a Secondary Mathematics Teacher Preparation Program*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Natural and Health Sciences in School of Mathematical Sciences, Program of Educational Mathematics

Accepted by the Doctoral Committee

____
Robert A. Powers, Ed.D., Chair

____
Dean E. Allison, Ph.D., Committee Member

____
William L. Blubaugh, Ph.D., Committee Member

____
Jennifer L. Harding-DeKam, Ed.D., Faculty Representative

Date of Dissertation Defense____

Accepted by the Graduate School

____
Robbyn R. Wacker, Ph.D.
Assistance Vice President of Research and Extended Studies
Dean of the Graduate School

ABSTRACT

Champion, Joseph Keith. The Mathematics Self-Efficacy and Calibration of Students in a Secondary Mathematics Teacher Preparation Program. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2010.

Social cognitive research has linked students' perceived academic capabilities, or self-efficacy, to academic choices, self-regulation, and performance in diverse contexts from reading comprehension to mathematical problem solving. This study addressed a need to investigate the interactions among prior achievement, self-efficacy, calibration (the accuracy of self-efficacy beliefs), and mathematics performance for students enrolled in the content courses of a secondary mathematics teaching program. The sample included 195 students in 12 classes ranging from calculus to second-semester abstract algebra at a mid-sized U.S. doctoral-granting university with a large secondary mathematics teacher education program. Data included background surveys, self-efficacy ratings preceding final exams, completed final exams, and transcripts of interviews with 10 secondary mathematics majors. Data analysis utilized structural equation modeling, analysis of variance, and thematic coding. Findings from both quantitative and qualitative analyses suggested participants' perceptions of their prior math performance, together with strong self-efficacy and slight overconfidence, were most associated with increased final exam performance. The discussion includes potential implications of the study for the content preparation of secondary mathematics teachers.

Keywords: self-efficacy, calibration, undergraduate mathematics, preservice mathematics teachers, structural equation modeling, social cognitive theory

ACKNOWLEDGEMENTS

Since this is an educational study, I would like to acknowledge some inspirational teachers. Melissa Champion, the physical education teacher, means more to me than anything, and her patience and joy inspire me daily. My research mentor, Rob Powers, taught me to think about what really matters – both in research and in life – and I hope to keep relearning that lesson throughout my life. I learned responsibility and hard work from my parents, Keith and Kathy, and I thank my siblings, Sharon, Laura, Tony, Ken, Marissa, and Brent, for my strange personality. Joel Iiams' soft-sell tactics first convinced me mathematics could be fun. Rachael and Sean Welder showed me a caring marriage, which was a critical part of my education. I thank the faculty at UNC for helping me to think of math education as respectful and exciting, especially Tensia Soto-Johnson, Jodie Novak, Anton Dzhamay, Steph Fitchett, and Michelle Chamberlain. Frieda Parker and Frank Pajares are my role models, and we all need role models.

Most directly related to my dissertation, I would like to thank Susan Hutchinson for her expert insights on structural equation modeling and my committee members, Jenni Harding-DeKam, Bill Blubaugh, and Dean Allison for their time and interest. John Fox's inclusion of the sem package in *R* saved and inspired untold hours of coding. My graduate student colleagues, especially Ann Wheeler, Bernadette Mendoza-Brady, Kristin King, Coralle Haley, Dave Glassmeyer, and Lee Roberson helped, encouraged, and laughed at me when I needed it. Finally, I want to thank the hundreds of students who keep me excited about what I do and who have taught me to ask for help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

"Whether you think that you can or that you can't, you're usually right" (Moncur, 2007, p. 1). When it comes to learning and teaching mathematics, the preceding statement, attributed to inventor Henry Ford, suggests a two-way relationship between belief in one's ability to complete a mathematical task and subsequent performance in the task. The social cognitive view of learning refers to this potential relationship as *perceived self-efficacy*, or self-evaluations of one's ability to accomplish given performances under specific constraints (Bandura, 1997). In the context of the mathematics completed by prospective secondary mathematics teachers, Ford's seemingly simple axiom about the influence of self-efficacy on success is just one part of the multifaceted relationships between self-beliefs, academic motivation, and performance that can affect students' experience of mathematics.

Research into self-efficacy has established that learners who express high self-efficacy in an academic domain tend to perform better on tasks in the domain than peers who report low self-efficacy (Pajares & Schunk, 2001). More than 1,800 research studies in education have addressed self-efficacy, and results suggest moderate-to-strong positive effects of self-efficacy judgments on performance tasks in domains as diverse as reading comprehension, career choice, and problem-solving in mathematics (Lightsey, 1999). However, there are documented exceptions to this trend (e.g., Klassen, 2006), and some important aspects of mathematics self-efficacy, especially *calibration*, or the accuracy of students' self-efficacy judgments, have received relatively little research attention.

The dissertation study reported here addressed mathematics self-efficacy by incorporating three aspects of self-efficacy identified in the literature as areas for future research: (1) the self-efficacy of college students in advanced mathematics courses, (2) the calibration of students' beliefs in their mathematical abilities, and (3) the mathematics self-efficacy of prospective secondary mathematics teachers. The guiding research question was: How do self-efficacy and calibration influence the exam performance of students enrolled in the advanced mathematics courses of a secondary teacher preparation program at a mid-sized liberal arts university?

The first sections in this chapter outline the research problem, purpose, and conceptual framework informing the study. Then, the narrative describes two pilot studies which provided preliminary findings for the guiding research question in college algebra and calculus settings. An overview of the research design is presented, including research questions and hypotheses, followed by a discussion of the significance of the study in terms of research, theory, and practice. Subsequent chapters include the review of literature (Chapter II), a description of the study methodology (Chapter III), quantitative and qualitative results (Chapter IV), and a synthesis and discussion of the findings in the context of related literature and potential follow-up studies (Chapter V).

Research Problem

In 2008, there were approximately 128,500 secondary teachers of mathematics in the United States, the vast majority of which (87%) teach exclusively mathematics (Morton, Peltola, Hurwitz, Orlofsky, Strizek, & Gruber, 2008). What do these teachers need to know and be able to do? On the federal level, the No Child Left Behind Act [NCLB] (2001) mandates all teachers earn “highly qualified” status, which requires teachers to (1) fulfill state certification requirements, (2) obtain at least a bachelor's degree, and (3) demonstrate “subject matter expertise.” To many in mathematics education, subject matter expertise is seen as the development of teachers’ mathematical content knowledge and pedagogical content knowledge (Hill, Rowland, & Ball, 2005). *Mathematical content knowledge* includes knowledge of and about mathematics and dispositions toward mathematics (Kahan, Cooper, & Bethea, 2003), while *pedagogical content knowledge* refers to understandings of mathematics that are particularly useful for teaching mathematics (Shulman, 1986).

Prospective secondary mathematics teachers build the content knowledge they need as teachers in large part through university mathematics coursework required for a bachelor’s degree in mathematics (Philippou & Christou, 1998). Such coursework can include topics such as calculus, differential equations, linear algebra, real analysis, geometry, and abstract algebra, and is hereafter collectively referred to as *advanced mathematics*. Monk’s (1994) survey of the content preparation of secondary mathematics teachers found participants completed a mean of 7.7 (*SD* = 4.3) advanced mathematics courses and only a mean of 1.9 (*SD* = 2.3) mathematics education courses in college. However, advanced mathematics coursework does not necessarily translate to “effective teaching” (Kahan et al., 2003) or strong pedagogical content knowledge (Hill et al., 2005), and possibilities for mathematics coursework to influence a future teacher’s practices may be substantively influenced by the teachers’ perceptions of their mathematical abilities.

Monk’s (1994) analysis of mathematics teachers’ content preparation suggests the positive effects of taking additional mathematics courses on student achievement diminish after about five courses, and a teachers’ completion of advanced mathematics coursework in college had only a small positive effect on student performance in advanced secondary mathematics courses such as calculus and had no statistical effect on student performance in remedial mathematics courses. That is, there are research indications of a somewhat tenuous connection between completing advanced mathematics coursework as a prospective teacher and developing the knowledge needed for teaching mathematics.

One consideration in teacher preparation has been inquiry into prospective teachers’ beliefs and attitudes about mathematics (Harding-DeKam, 2005) within the context of their preparation in advanced mathematics. As Philippou and Christou (1998) point out, “teachers' formative experiences in mathematics emerge as key players in the process of teaching since what they do in the classroom reflects their own thoughts and beliefs” (p. 191). In particular, Thompson’s (1984) inquiry into teachers’ beliefs found self-beliefs and perceptions of mathematics coursework work in concert with beliefs about the discipline of mathematics to influence teachers’ instructional choices, and, ultimately, to impact student achievement. However, scarce research has addressed prospective secondary teachers’ perceptions of their own mathematical capabilities, especially in the context of advanced mathematics courses and research is needed to investigate prospective teachers’ mathematics self-efficacy toward successfully completing advanced mathematics.

Purpose Statement

The purpose of this concurrent mixed methods study (Creswell, 2003) was to examine relationships among the strength and accuracy of mathematics self-efficacy beliefs and the subsequent performance of students enrolled in advanced mathematics courses required by the secondary mathematics education program at a mid-sized university in the Rocky Mountain West. Utilizing constructs and hypothesized relationships from social cognitive theory, broad statistical relationships derived from in-class survey and assessment data were supported by task-based interviews to address the research problem through seven research questions.

The quantitative purpose of the study was to estimate effects of participants' self-efficacy and calibration on subsequent mathematics performance using a social cognitive model for performance in advanced mathematics. Intervening variables included the difficulty of exam tasks, the amount of required mathematics in participants' chosen college majors, participants' gender, and indicators of participants' high school mathematics achievement. A parallel qualitative strand of the investigation explored mathematics self-efficacy and calibration through the rich information provided by task-based interviews. The quantitative and qualitative strands then converged to contrast, triangulate, and validate findings and provide insights which may not have been possible through an exclusive reliance on either strand.

Conceptual Framework

Social cognitive theory provides a foundational framework for considering prospective mathematics teachers' self-beliefs of their mathematical capabilities. When considered in the complicated context of advanced mathematics content preparation, a social cognitive framework can help explain the level and accuracy of prospective teachers' self-perceptions of their abilities in the mathematics courses required by secondary mathematics teacher preparation programs.

Overview of Social Cognitive Theory and Self-Efficacy

Albert Bandura's social cognitive theory first began as a means for explaining observational learning mechanisms by positing that a *causal triadic reciprocality* exists between individuals' behavior, environmental stimuli, and internal cognitive factors (Simon, 1999). This approach has since developed into a robust theory increasingly focused on the cognitive and motivational processes supporting metacognition (Schraw, 1998), self-efficacy, and self-regulation among learners as they acquire knowledge and skills (Martin, 2004). In particular, perceived *self-efficacy,* or judgments of one's ability to accomplish given performances in particular contexts (Bandura, 1997), is a particular focus of social cognitive research in mathematics education. Lightsey (1999) identified over 2500 hundred articles addressing positive relationships between self-efficacy and achievement.

Social cognitive research considers self-efficacy to be a primary mediating mechanism in all human cognition because self-beliefs in ability act as a filter between prior experiences and subsequent development of abilities within a particular domain. In contrast to self-concept, which refers to more global self-beliefs and personal identity, Pajares and Schunk (2001) summarize the hypothesized direct role self-efficacy plays in the choices people make:

> Self-efficacy beliefs influence the choices people make and the courses of action they pursue. Individuals tend to engage in tasks about which they feel competent and confident and avoid those in which they do not. Efficacy beliefs also help determine how much effort people will expend on an activity, how long they will

> persevere when confronting obstacles, and how resilient they will be in the face of adverse situations. (p. 241)

Attributed in part to individuals' tendencies to rely heavily on self-efficacy beliefs during difficult tasks (Bandura, 1997), self-efficacy judgments are often better statistical predictors of performance in academic domains than standardized measures of ability or intelligence (Pajares & Kranzler, 1995). In fact, after controlling for instructional factors, path analyses of performance incorporating biographical (e.g., socio-economic status, gender), motivational, and instructional variables, suggest self-efficacy beliefs account for the largest portion of variation in academic performance (Madewell & Shaughnessy, 2003). Though measures of self-efficacy are often useful for predicting performance, there is evidence that strong self-efficacy beliefs themselves do not guarantee success in difficult domains such as mathematics. In particular, developing both strong and accurate self-efficacy beliefs may be the key to self-efficacy's benefits in learning mathematics.

Mathematics Self-Efficacy, and Calibration

Underscoring the complex nature of students' confidence in their mathematical abilities and performance on closely matched mathematical tasks, Chen and Zimmerman (2007) found that U.S. seventh graders reported much higher mathematics self-efficacy beliefs than sixth grade Taiwanese students, yet the U.S. students performed significantly worse than the Taiwanese students on corresponding mathematics tasks. That is, the U.S. students displayed a larger tendency toward overconfidence in their self-efficacy ratings than the tendency toward more accurate self-efficacy ratings among Taiwanese students. Linking academic behaviors such as reduced effort to overconfidence, Chen and Zimmerman suggest the cross-cultural differences in overconfidence may contribute to larger trends toward underperformance by U.S. students in mathematics when compared to Taiwanese students.

Sometimes referred to as “feeling-of-knowing accuracy” (Schraw, 1995, p. 326), students’ *calibration* (Pajares & Miller, 1994) in self-efficacy ratings is a relatively new area for research in mathematics education with foundations in experimental psychology and reading education (Lin & Zabrucky, 1998). The tendency of students across educational levels and performance abilities toward *overconfidence*, or positively biased judgments (Schraw), has been reported in studies of college students’ self-efficacy for reading tasks, in particular. In their review of literature addressing the calibration of adult readers, Lin and Zabrucky refer to this tendency as an “illusion of knowing” effect and suggest possible detrimental effects of overconfidence:

> There is a tendency for adult students to generate unrealistic feelings of knowing when it comes to evaluating outcomes of learning. As can be seen in the present review, overconfidence is a common phenomenon among young adult students that may result in inadequate learning due to premature termination of cognitive processing. (p. 384)

Bandura (1997) suggests slight overconfidence in one’s self-efficacy can be psychologically adaptive because overconfidence can have positive benefits on effort and persistence. In this view, poor calibration in the form of overconfidence can be reframed as a set of optimistic self-evaluations that may ultimately support taking-on challenges. Nonetheless, Bandura and other calibration researchers (e.g., Pajares & Kranzler, 1995) caution against grossly inflated overconfidence, suggesting that unrealistic overconfidence can lead students to engage in self-handicapping academic behaviors (Urdan, 2004) such as reduced studying and increased procrastination.

From a quantitative perspective, there is support for calibration as a measure that contributes to statistical explanations of variation in achievement beyond the variation explained by self-efficacy judgments and prior achievement in mathematics (Pajares & Miller, 1997). Chen (2003) found U.S. middle school students at every ability level tend to show poor calibration in the form of overconfidence, but also that self-efficacy and calibration provide significant and independent predictive value in a path analysis model for mathematics performance.

One hypothesis regarding calibration is that learners may grow to be more accurate in assessing their abilities through a content-specific developmental process (O'Connor, 1989). In a review of calibration research from experimental psychology in the 1960s to 1980s, O'Connor identified several factors influencing calibration: (1) familiarity with task requirements (e.g., assigning numbers to feelings of uncertainty), (2) familiarity with the topic of interest (subject matter knowledge), and (3) feedback on the accuracy of prior judgments. O'Connor also describes research that college students' self-efficacy to attain final letter grades in their courses tends to be well-calibrated, suggesting students may develop good calibration in predicting general academic outcomes while simultaneously demonstrating poor calibration in their self-efficacy to complete specific course-related tasks.

Through mathematics self-efficacy and calibration, social cognitive theory provides a foundation for interpreting the mathematical confidence and achievement of prospective secondary mathematics teachers in advanced mathematics. However, social cognitive theorists do not subscribe to global models of self-efficacy and performance (Bandura, 1997), because personal, social and cultural conditions are seen as important co-determinants of academic confidence, motivation, and behaviors. Thus, it was important to develop a hypothesized model of self-efficacy, calibration, and performance based on the specific learning context at the research site. Two pilot studies informed this effort.

Two Pilot Studies

In preparation for this study, the researcher conducted two pilot studies at the research site; the first study focused on the predictive value of mathematics self-efficacy and calibration in College Algebra *(N* = 128) during Fall 2007, and the second study extended and refined the methodology of the first study in the context of Calculus I (*N* = 119) during Spring 2008. The first pilot study was set within a larger study of student achievement and goal structures that incorporated balanced, random assignment of students to two instructional conditions, one of which included a classroom communication system featuring a network of graphing calculators and a classroom presentation system. Within the college algebra study, the first pilot study used a concurrent mixed methods (Creswell, 2003) design to investigate students' self-efficacy ratings, calibration, and experiences of course feedback in the four college algebra sections throughout the semester. The second pilot study utilized a post-test only with non-equivalent groups design (Creswell, 2003) to further validate and refine the measures and statistical model for the effects of self-efficacy and calibration on final exam performance in the population of students enrolled in advanced mathematics.

Quantitative results from the first pilot study confirmed many of the self-efficacy research findings that had previously been attributed to middle and secondary school students (e.g., Chen, 2003). The survey techniques used in the study mirrored procedures used in earlier social cognitive studies of calibration (e.g., Chen, 2002; Pajares & Miller, 1994) and incorporated two measures of calibration—*accuracy*, which is an absolute measure, and *bias*, which is a directional measure of calibration—in part to compare the predictive utility of each measure. Self-efficacy, accuracy, and performance scores were converted to a five-point ordinal scale (i.e., 0 = lowest, 5 = highest), and calibration was expressed on a 10-point ordinal scale (e.g., -5 = underconfident, 0 = calibrated, +5 = overconfident). Descriptive statistics for the four measures are shown in Table 1 and suggest participating college algebra students tended to express self accuracy ratings which were moderately accurate, but consistently overconfident. Correlation analysis of the variables confirmed findings from Chen and Zimmerman (2007) that self-efficacy, mathematics performance, and calibration bias and accuracy are all significantly intercorrelated at the $\alpha = 0.01$ criterion (see Table 2).

Table 1.

Means and Standard Deviations of Measures in the First Pilot Study

(N = 91)	Performance		Self-Efficacy		Bias		Accuracy	
Measure	M	SD	M	SD	M	SD	M	SD
Exam 1	4.11	0.62	4.00	0.69	1.07	0.96	2.82	0.67
Exam 2	3.01	0.89	3.54	0.89	1.67	1.08	2.46	0.66
Exam 3	3.70	0.74	3.62	0.84	0.98	1.04	2.82	0.75
Final Exam	4.01	0.75	3.84	0.80	1.07	1.18	2.77	0.78
Combined	3.73	0.55	3.76	0.70	1.18	0.72	2.73	0.49

Table 2.

Correlations for Composite Measures in the First Pilot Study

	Bias	Accuracy	Self-Efficacy	Performance
Bias	–	-0.60	0.36	-0.51
Accuracy		–	0.36	0.76
Self-Efficacy			–	0.54
Performance				–

Data analysis in the first pilot study led to a multiple linear regression model which included composite measures of self-efficacy and calibration bias as predictors of students' performance on four in-class examinations. Analysis of the model suggested the data met the four assumptions of linear regression modeling (Osborne & Waters, 2002), including (1) linear relationships between the independent and dependent variables, (2) independence of errors, (3) normality of variables, and (4) equal variances in errors (homoscedasticity). The regression model was significant ($F = 265.4$, $p < 0.001$) and yielded an R^2 value of .86 ($SE = .2$), suggesting 86% of variance in college algebra students' performance could be explained by independent linear effects due to calibration and self-efficacy. Standardized regression coefficients showed strong and approximately equal effects of bias ($\beta = -.82$) and self-efficacy ($\beta = .83$). That is, while increasing self-efficacy judgments was associated with increased mathematics performance, tendencies toward overconfidence were approximately equally associated with decreased performance among the college algebra study participants.

The qualitative inquiry component of the first pilot study looked at college algebra students' experiences of multiple sources of feedback in technology-enriched instructional settings, including online homework, graphing calculators, course management software, and small-group activities, in four sections taught by the two instructors, of which one was the researcher. One section taught by each instructor utilized a classroom graphing calculator network as a communication and presentation system during class activities. Using purposeful sampling techniques (Glesne, 2006) in conjunction with suggestions from the class instructor, the qualitative investigation included data from interviews of seven students and digital artifact analyses (e.g., saved computer screenshots) as part of a holistic comparative case study (Merriam, 1998) of students experiences in the two instructional settings.

Results from the qualitative strand of the first pilot study suggested students relied heavily on performance feedback and mastery experiences in the form of quizzes, exams, and online homework as well as social comparisons to classroom peers in forming self-efficacy evaluations of their content understanding. These preliminary qualitative findings pointed to considering Bandura’s (1997) *four sources of self-efficacy*—mastery experiences, social persuasions, vicarious experiences, and physical and emotional states—as a potential qualitative framework for exploring the relationships among self-efficacy, calibration, and performance for students enrolled in advanced mathematics courses.

The Calculus I pilot study yielded similar results to the quantitative strand of the college algebra pilot study regarding the correlations and predictive value of self-efficacy and calibration toward students’ exam performance. This second pilot research design collected less data from each student (a single exam versus four) and was less controlled than the first pilot study because the cross-sectional design did not include random assignment of students to sections and the participating calculus instructors used different exams and self-efficacy instruments. In addition to a decrease in statistical power due to reduced sample size (Frankfort-Nachmias & Nachmias, 2000), it was expected that any linear regression models would account for a lower proportion of variation in students' mathematics performance. However, as in the first pilot study, self-efficacy and calibration bias accounted for large independent portions of the variation in exam performance, collectively explaining R^2 = 76% of the variation in calculus students' final exam performance. However, both students' performance and calibration bias on final exams varied greatly across course sections, which suggested future research might follow Chen's (2003) consideration of potential differences in self-efficacy and calibration by the level of difficulty in test items.

While the two pilot studies suggested some relationships between calibration, self-efficacy, and mathematics performance in advanced coursework, interpretation of the data analysis was limited by an assumption in multiple linear regression that independent variables do not include measurement error (Frankfort-Nachmais & Nachmias, 2000). Although observed reliability coefficients of self-efficacy and calibration measures are typically moderate to strong (O'Connor, 1989; Pajares & Miller, 1997), the fact that each measure includes self-reports of latent psychological variables suggests that structural equation modeling is more appropriate, especially in light of the strong theoretical support for directional relationships among calibration, self-efficacy, and mathematics performance (Pajares & Kranzler, 1995).

In summary, the pilot studies informed the research design in four important ways. First, the procedures in the pilot studies helped refine the data collection protocol and helped to establish the feasibility of the data collection and analysis procedures at the research site. Second, the regression findings from both studies suggested strong, approximately equal, and opposite effects of self-efficacy and calibration on performance. Third, the qualitative inquiry pointed to Bandura's (1997) conception of the four sources of self-efficacy as a conceptual tool for investigating relationships between self-efficacy, calibration, and performance. Finally, methodological considerations suggested the appropriateness of using structural equation modeling in future mathematics self-efficacy research.

Hypothesized Model and Research Questions

A central purpose of the research was to address the research problem by investigating a social cognitive model for advanced mathematics performance that incorporated self-efficacy, calibration, and the amount of mathematics in students' major as endogenous variables and high school mathematics achievement as a single exogenous variable. The model, shown in Figure 1, was based on an extensive review of related literature and was similar to models used by Chen (2003) and Pajares and Kranzler (1995) in studies of mathematics self-efficacy among general student populations.

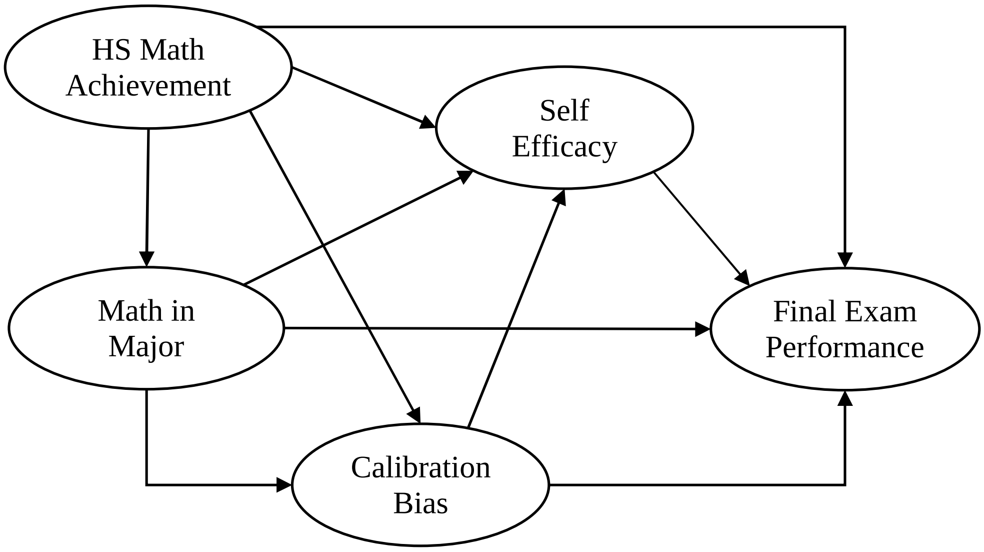

Figure 1. Hypothesized path model for performance in advanced mathematics.

One distinguishing characteristic in the structural model was the inclusion of hypothesized effects of the amount of mathematics in students' college majors as having a potential influence on self-efficacy, calibration and performance among students enrolled in advanced college mathematics courses. The hypothesized model is a compact way of representing four quantitative research questions (Q1-Q4 below) that were addressed using structural equation modeling. In addition, two quantitative questions addressed potential differences in the endogenous variables by the intervening variables corresponding to students' gender (Q5) and the difficulty of exam items (Q6), each of which were addressed through multivariate analysis of variance (MANOVA) procedures. Finally, a single qualitative research question called for a holistic description of the processes relating self-efficacy, calibration, and mathematics performance for the important subpopulation of prospective secondary mathematics teachers.

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects on participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

Q7 In what ways do prospective secondary mathematics teachers' mathematical problem-solving compare and contrast with the hypothesized relationships between self-efficacy, calibration, and performance in advanced mathematics?

A primary purpose of the literature review (Chapter II) was to ground the research questions within social cognitive theory and related literature on mathematics self-efficacy. In addition, the review of literature provided the rationale for directional effects in the structural model and led to the development of hypotheses (listed at the end of Chapter II) to correspond to each of the research questions.

Brief Overview of the Research Design

The research design incorporated a social cognitive perspective on cognition and academic achievement that emphasized the mediating roles of self-efficacy and calibration on students' performance in mathematics. The methodology used a concurrent triangulation strategy for mixed-methods (Creswell, 2003), including a qualitative inquiry to cross-validate and contextualize findings from a statistical model of students' self-efficacy, calibration, and performance in the content courses of a secondary mathematics teacher education program at a single mid-sized (enrollment of about 12,000) liberal arts university in the Rocky Mountain West. Data collection included quantitative self-efficacy surveys and exam performance scores for a sample of 195 students in 12 advanced mathematics classes along with qualitative task-based interview responses from 10 purposefully sampled participants. Details of the methodology appear in Chapter III.

Dissemination of Findings

The study findings were disseminated in three ways. First, this dissertation narrative was completed as part of the researchers' doctoral degree requirements and made available to the public through the University of Northern Colorado's library system. Second, the study and findings were summarized in a professional research presentation at a national conference on mathematics education and through research presentations in five U.S. mathematics departments that specialize in the preparation of secondary mathematics teachers. Finally, the researcher expects to synthesize the study and findings into a scholarly article and to submit the article to a peer-reviewed mathematics education journal. The intended audience of the dissertation and research presentations was primarily faculty responsible for preparing future secondary mathematics teachers, including mathematics professors and teacher educators, but also included educational psychologists, educational researchers, secondary mathematics majors, and those interested in the self-beliefs of students in advanced mathematics courses.

Significance of the Study

The aim of this section is to outline some anticipated implications of the study for research, theory, and the preparation of secondary mathematics teachers; for a more detailed discussion of the significance of study in the context of study limitations, see Chapter V. Based on the review of literature and pilot studies, the research study was expected to (1) add to existing self-efficacy research by including an important and often overlooked population of participants, (2) partially fill a need for mixed methods studies in social cognitive research, (3) add to research on the mathematical content knowledge and self-beliefs of prospective mathematics teachers, and (4) inform the practice of the mathematical content preparation of prospective secondary mathematics teachers.

First, the review of literature identified substantial needs for research addressing the self-efficacy and calibration of college students. The research design could lead to findings regarding the value of using these measures to predict student performance in advanced mathematics, as well as describe potential intervening effects of students' prior achievement, gender, and college major. Moreover, the qualitative inquiry could suggest new quantitative avenues for evaluating the generalizability of themes emerging from the exploratory task-based interviews.

Second, social cognitive theory posits a dynamic interplay between learners' perceptions of their performance, self-assessments of capability, and academic choices (Pajares & Urdan, 2006), and this approach to learning necessarily admits the effects of rich constellations of context informed by life experience and culture (Bandura, 1997). However, nearly all existing mathematics self-efficacy research has employed quantitative methods (Usher & Pajares, 2008). This study, by blending quantitative and qualitative techniques, was expected to help describe the context and processes through which self-efficacy and calibration influence performance among prospective secondary mathematics teachers. This mixed methods approach allowed both statistical testing of broad-scale effects and emergent inquiry into mathematics self-efficacy and calibration among preservice secondary mathematics teachers.

Third, the research design had the potential to build on emerging understandings of social learning in mathematics as it relates to the practice of secondary mathematics teacher preparation. Future teachers need to know what mathematics they understand well (Ball & McDiarmid, 1989), and the study findings could help describe the qualities of, and processes supporting, the metacognitive aspects of mathematics learning related to self-efficacy and calibration among students taking advanced mathematics courses. These descriptions, by including the important population of prospective secondary mathematics teachers, can buttress efforts to prepare high school mathematics teachers that are realistically confident in their mathematical skills.

Finally, the research design had the potential to help inform educational interventions to promote adaptive mathematics self-efficacy in the content courses of secondary mathematics teacher preparation programs. Citing evidence of overconfidence in students at every educational level, Pajares and Miller (1997) highlight the significance of developing a better understanding of calibration in mathematics students because of the import of affecting students' calibration:

> It may be more important to develop instructional techniques and intervention strategies to improve students' calibration than to attempt to raise their already overconfident beliefs. Improved calibration should result in better understanding by students of what they know and do not know so that they more effectively deploy appropriate cognitive strategies during the problem-solving process. (p. 216)

The quantitative and qualitative findings describing mathematical self-efficacy and calibration among secondary mathematics majors may help to suggest ways in which teacher educators and mathematics professors can set conditions in which students can develop robust and realistic perceptions of their mathematics competencies.

CHAPTER II

REVIEW OF LITERATURE

Imagine two friends, Casey and Jesse, preparing for a final exam in their calculus course. Throughout the semester, the study-buddies met to do homework a few nights a week, prepared together for exams, and experienced similar high marks on exams and graded assignments. Encouraged by her success in this and other mathematics courses, Casey is looking forward to the final exam. She expects to do about as well on the final exam as she did on the midterm exams, plans to study alone by rereading her notes, and will go to the final exam feeling calm and confident. Jesse, however, is concerned about the exam. Jesse tells her friend Casey that she is always worried about making "stupid mistakes" on exams, and she is worried that she may have forgotten much of the content from early in the semester. Besides, without Casey to help her study, Jesse does not like her chances of doing well on the exam.

The hypothetical situation of Casey and Jesse just before the final exam raises some questions that can be partially answered by research into the interplay between academic experiences, self-efficacy, and performance. Will Casey's self-assuredness in her mathematics abilities be likely to help or hinder her when it comes to her performance on the final exam? Do students who, like Jesse, have lower self-efficacy in an advanced mathematics course, become discouraged and study less than their more confident peers, or do they find ways to overcome their concerns to ultimately achieve higher levels of performance? To what extent might Jesse and Casey's performance in prior mathematics classes, their gender, or even the difficulty of their upcoming calculus exam influence their self-efficacy?

This chapter describes a base of scholarly literature and conceptual framework on which the dissertation study rests. The first sections detail a theoretical foundation for approaching mathematics learning through concepts in social cognitive theory, including self-efficacy and the accuracy of confidence judgments. Next, the narrative narrows to research describing social cognitive views of mathematics performance, including empirical and theoretical models for mathematics performance that incorporate self-efficacy and related motivational variables. This is followed by rationale for the hypothesized model for advanced mathematics performance used in the research. With the base of scholarly research supporting the variables and theoretical perspective, the review of literature culminates in research questions and hypotheses.

Overview of Social Cognitive Theory

Social cognitive theory originated in the neo-behaviorist research program of Albert Bandura in the 1950s and 1960s (Schunk, 2004), which included the classic 1961 Bobo Doll experiment at Stanford University. The Bobo Doll experiment traced increases in aggressive behaviors in preschool children to observing peers or cartoons displaying similar behaviors on film (Bandura, Ross, & Ross, 1963). Bandura's experiments gave evidence for learned aggressive behaviors in conditions that contained no observable reinforcements. These and related results contrasted sharply with Skinner's operant conditioning learning theory, which was the dominant learning theory at the time in psychology (Pajares & Schunk, 2001). To help explain the Bobo Doll experiment findings, Bandura developed a social learning theory that emphasized *observational (or vicarious) learning* through behavioral and cognitive modeling.

Embracing the fact that much human learning occurs in social contexts, social cognitive researchers initially focused on processes that link observed behaviors, social comparisons, and personal motivation, such as response facilitation (going along with the crowd), inhibition (observing others being punished), disinhibition (observing others not being punished), and the attention, retention and production of modeled skills (Schunk, 2004). Research results suggested observational learning can be influenced by (1) intellectual and physical development, (2) the perceived prestige of models, (3) vicarious experience of the consequences of modeled behaviors, (4) personal goals and outcome expectations, and (5) perceived self-efficacy in a given domain (Schunk, 2004).

Following the emphasis on observational learning in the 1960s and 70s, social cognitive theory evolved to incorporate principles in the social constructivism paradigm (Simon, 1999), and grew to focus on the causal processes underlying the effects of self-beliefs on behavior (Bandura, 1995). In the modern social cognitive view of learning as an *agentic* process, people rely on self-perceptions to choose actions that exert influence and establish control over their environment. This agency results in a *triadic reciprocality* between personal factors (i.e., cognitive, affective, and biological), behaviors, and external stimuli (Schunk, 2004). Theorists consider three broad types of personal cognition to have mediating effects on the reciprocal nature of social learning: self-efficacy, self-regulation (e.g., Zimmerman & Schunk, 1989), and outcome expectancies (Bandura, 1997). Of the three mediating constructs, the study focused on self-efficacy, which Bandura (1997) cites as having the strongest mediating effect on learning. The following section defines self-efficacy and places it in the context of related constructs in educational psychology.

Self-Efficacy and Related Constructs

Definition of Self-Efficacy

Bandura (1997) defined *perceived self-efficacy* as “beliefs in one’s capabilities to organize and execute the courses of action required to produce given attainments” (p. 3). Together with the social cognitive view of knowledge as personally and socially constructed within cultural milieus (Simon, 1999), perceived self-efficacy can include self-confidence in one’s ability to exercise control in a variety of circumstances, such as self-efficacy to regulate affective (emotional) states, to change social conditions, or to achieve a desired performance level on a mathematics test. According to Bandura (1997):

> [Self-efficacy] beliefs influence the courses of action people choose to pursue, how much effort they put forth in given endeavors, how long they will persevere in the face of obstacles and failures, their resilience to adversity, whether their thought patterns are self-hindering or self-aiding, how much stress and depression they experience in coping with taxing environmental demands, and the level of accomplishments they realize. (p. 3)

Thus, self-efficacy is primarily important to educational researchers because of the effects of self-efficacy on students’ choices, motivation, and persistence (Bouffard-Bouchard, 1990). Although Bandura’s preceding quote seems to ascribe a kind of universality to the influence of self-efficacy on one’s life, the meaning and role of self-efficacy in learning can be better operationalized by considering related constructs in the theory of academic motivation.

Self-Efficacy in the Context of Academic Motivation

Educational psychologists who focus on academic motivation point to a constellation of *self-beliefs*, or set of conceptions one has about oneself (Pajares &

Schunk, 2002), that combine in complex ways to influence learning. Thus, one way to operationalize the meaning of self-efficacy in mathematics learning is to consider self-efficacy in the context of other related self-belief constructs. Table 3 includes examples of self-belief statements in mathematics that typify some constructs in academic motivation literature that have similarities with self-efficacy, including self-concept of ability, self-esteem, outcome expectancies, locus of control, affective confidence, goal orientations, and social comparisons.

Table 3.

Examples of Self-Belief Statements Related to Mathematics

Statement	Motivational Construct
I can solve this quadratic equation.	Self-efficacy
I can earn at least a C in mathematics class this semester.	Self-efficacy
I am really good at graphing functions.	Self-concept of ability
I am smart in mathematics.	Self-Esteem
I will earn at least a B on my algebra test tomorrow.	Outcome expectancy
When I work hard, I tend to do well in mathematics.	Locus of control (internal)
My teacher will pass me if I turn in all my homework.	Locus of control (external)
I feel like I am ready to learn the quadratic formula.	Affective confidence
I want to earn at least a C on my algebra test tomorrow.	Performance goal
I want to understand function notation.	Mastery goal
I am better than my friends are at doing mathematics.	Social comparison

One of the oldest terms in modern psychology is *self-esteem*, which was defined by William James (1890) in the 19th century as feelings of self-worth that arise from accomplishing some fraction of what one wishes to accomplish. Decades later, a trend emerged out of the humanistic movement in psychology during the 1960s through 1980s (Pajares & Schunk, 2001) that emphasized promoting feelings of self-esteem in school children with the hopes of providing a foundation for academic success. The self-esteem movement was challenged by policy critiques that self-esteem programs in schools promoted egocentrism and excessive praise (McMillan, Singh, & Simonetta, 2001). Among the hundreds of research studies into the link between self-esteem and performance, most reported weak associations between self-esteem and performance (Hansford & Hattie 1982). In their extensive review, Hansford and Hattie found the mean reported correlation between self-esteem and academic performance in 128 studies at the K-12 level to be $r = 0.21$, with reported values ranging from $r = -.77$ to $r = .96$. The authors interpreted these results as providing less-than-overwhelming evidence for the value of self-esteem programs, and suggested some forms of self-esteem—what McMillan and colleagues refer to as unearned self-esteem—may actually have detrimental effects on academic functioning.

Widespread dissatisfaction with programs to promote self-esteem in schools (Pajares & Schunk, 2001) led to alternative formulations of what constitutes helpful self-perceptions in academics, particularly the constructs of self-concept, goal structures, locus of control, and self-efficacy. Among these, *self-concept* is most closely linked with self-esteem. Bong and Clark (1999) cite self-concept and self-efficacy as the two most researched self-related constructs in academic motivation theory, but also point to theoretical challenges when comparing self-efficacy and self-concept in academics.

While self-efficacy toward a given task is generally used to mean “the conviction that one can successfully execute the behavior required to produce the outcomes”

(Bandura, 1997, p. 79), self-concept has many operational meanings in motivation literature. Bong and Clark (1999) found self-concept most often refers to “one’s perception of the self that is continually reinforced by evaluative inferences and that it reflects both cognitive and affective responses” (p. 140). *Self-concept of ability*, refers to the cognitive component of self-concept, including descriptions (e.g., Can I do this task?) and self-evaluations (e.g., How well can I do this task?). In contrast, *self-esteem* encompasses the affective component of self-concept, and includes feelings of worth and approval or disapproval in a given learning domain. That is, self-esteem can be domain specific (e.g., self-esteem in mathematics may differ from self-esteem in reading) and can flow from one making what Moore and Small (2007) refer to as *normative comparisons* (e.g., Do I feel smarter after finishing my mathematics homework?) and *social comparisons* (e.g., Am I as good as my mother at doing mathematics?).

Some theorists restrict self-esteem to mean only general feelings of personal worth (e.g., Branden, 1994) and refer to related feelings that are specific to a domain as *affective confidence* (Reyes, 1984). In summarizing affective cognitive research in mathematics learning, Reyes identified consistent, though moderate, positive associations between affective confidence in mathematics and mathematics achievement. Reyes cites Dowling’s (1978) Mathematics Confidence Scale as a reliable measure of affective confidence in mathematics in college students, and since affective confidence is the component of self-concept that is easily conceptually differentiated from self-efficacy, Pajares and Miller (1994) used a modified version of Dowling’s survey as their measure of self-concept in undergraduate mathematics students.

Self-concept and self-efficacy both include cognitive perceptions of one's capabilities in a domain, but differ in generality and scope. A self-efficacy belief is restricted to "a judgment of one's capabilities to execute specific behaviors in specific circumstances" (Pajares & Miller, 1994, p. 194), while self-concept can include general assessments of ability, feelings of self-esteem, and inferences about one's performance in relation to their peers or perceived norms (Madewell & Shaughnessy, 2003). The more general scope of self-concept as a construct, along with the relative lack of consistent meaning of self-concept in research, is perhaps partly responsible for weaker and less consistent associations between self-concept and academic performance than those reported for self-efficacy and academic performance (Bong & Clark, 1999). In a path analysis of the mathematics performance of 350 undergraduate students in Georgia, Pajares and Miller (1994) found strong main effects of self-efficacy on performance and only moderate indirect effects of self-concept on performance. Moreover, participants' self-efficacy ratings showed higher internal consistency than self-concept ratings, and the influence of self-concept on performance was largely accounted for by a meditational influence of self-concept on self-efficacy.

While the preceding discussion situates self-efficacy in the context of self-esteem, self-concept, and affective confidence, some constructs in academic motivation theory encompass self-beliefs that do not directly reference capabilities, but nonetheless may influence academic success. In particular, intrinsic theories of motivation such as social cognitive theory view individuals' actions as proactive efforts to reach desired *personal or social goals*. Achievement Goal Theory (Alderman, 1999) describes two types of personal goal orientations that influence motivation and achievement: *mastery* *orientations* and *performance orientations*. Students with mastery orientations in a domain aim to learn because they believe learning in the domain is inherently valuable and meaningful, whereas students with performance orientations seek and evaluate levels of attainment by focusing on perceived normative or social standards (Urdan, 2004). While mastery goal orientations are typically associated with many of the same adaptive educational choices as self-efficacy, such as persistence, help-seeking, and taking-on new challenges (Urdan, Pajares, & Lapin, 1997), performance orientations have been associated with both adaptive and non-adaptive learning choices (Husman, Brem, and Duggan, 2005).

Goal theorists generally view *performance-avoidance* goals (i.e., to avoid performing below some given level) as having negative effects on academic functioning (Elliot & Moller, 2003), but the relative merits of *performance-approach* goals (i.e., aims to perform up to some desired level) are disputed. In the context of self-efficacy, Elliot and Moller's meta-analysis of performance-approach goals research identified a weakly positive effect of performance-approach orientations on students' academic achievement and self-efficacy, but found inconsistent results regarding the influence of performance-approach orientations on many other academic behaviors such as help-seeking and persistence. Midgley and Urdan (2001) found 7th grade mathematics students with performance orientations were more likely than students with mastery orientations to engage in *self-handicapping* behaviors, including avoiding studying and purposely not trying hard in mathematics classes. Likewise, students with performance-avoidance orientations engaged in more self-handicapping behaviors than those with performance-approach orientations.

Zimmerman, Bandura, and Martinez-Pons (1992) investigated the potential influences of self-efficacy on middle school social studies students' goal setting behaviors and academic performance. The authors' path analyses identified direct effects of self-efficacy on test performance, grades, and expressing confidence to set and attain short-range goals. However, potential relationships between self-efficacy, goal orientations, and achievement among college mathematics students have received little research attention (Elliot & Moller, 2003).

One notable exception to the lack of unification between goals research and self-efficacy research is the work of Lent, Brown, and Hackett (1994), who suggest that self-efficacy acts as an important mediator on the effects of prior achievement on outcome expectations, interests and goals, and future attainments. The authors' general model (see Figure 2), though originally intended for adolescent career-choice behaviors, is particularly useful for setting self-efficacy in the context of the personal, contextual, and experiential constructs that affect academic choices.

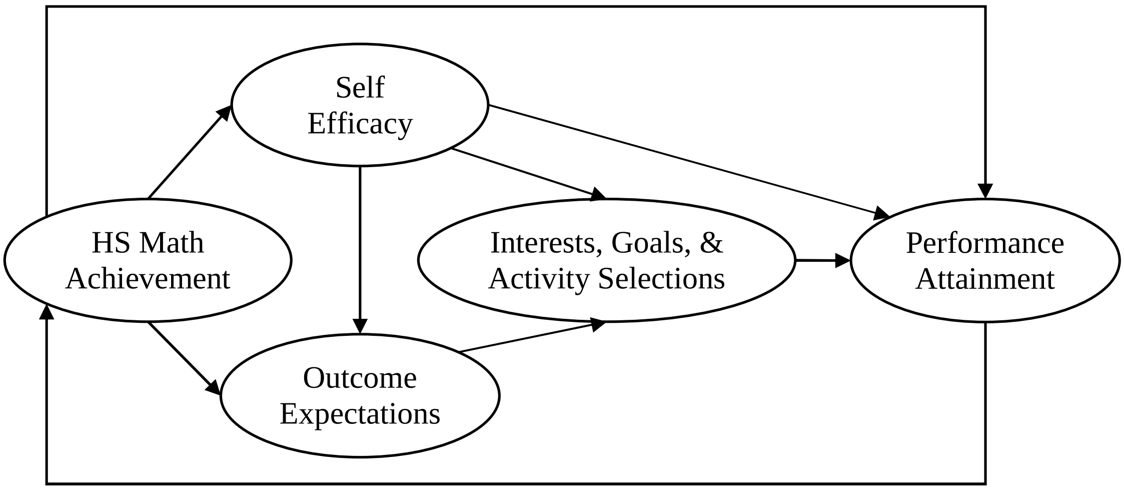

Figure 2. Lent and colleagues' model of career-choice behaviors (1994, p. 88).

Self-Efficacy in Mathematics Education

Beginning in the 1980s, studies of mathematics self-efficacy have linked students' self-efficacy to complete mathematical tasks to a variety of educational outcomes, including problem solving performance and persistence (Bouffard-Bouchard, 1990), choice of college major (Betz & Hackett, 1986), and career interests (Lapan, Shaughnessy, & Boggs, 1996). In fact, Pajares and Graham (1999) explain why social cognitive theorists have devoted so much study to describing relationships between mathematics self-efficacy and achievement:

> The area of mathematics has received special attention in self-efficacy research for a number of reasons. Mathematics holds a valued place in the academic curriculum; it is prominent on high-stakes measures of achievement generally used for level placement, for entrance into special programs, and for college admissions; and it has been called a ''critical filter'' for students in pursuit of scientific and technical careers at the college level. (p. 125)

Self-efficacy research has contributed potential explanations for some puzzling differences in students' motivation for and interests in mathematics and mathematics-related careers (Madewell & Shaughnessy, 2003), particularly in terms of differences in male and female students. For example, in their path analysis of the mathematics performance of 415 high school juniors, O'Brien, Kopala, and Martinez-Pons (1999) identified students' gender and mathematics self-efficacy as having direct effects on mathematical career interests, with mathematics self-efficacy in turn influenced by students' ethnic identity, prior academic achievement, and socio-economic status. In the post-secondary setting, Hackett and Betz (1989) found mathematics self-efficacy was a better predictor of choice of a mathematics-related college major among college students than indicators of either mathematics problem-solving performance or high school mathematics performance. The authors suggest men were more overconfident than were women in their self-efficacy for mathematical problem solving, but also suggest participants of both genders tend to be moderately overconfident in their problem solving skills.

A four-year longitudinal study by Lapan and colleagues (1996) looked at predictors of college major and interest in mathematics-related careers among 101 men and women from Grade 11 through the students' junior year of college. Findings suggested mathematics self-efficacy played an important mediating role on the effect of high school mathematics experiences on choice of college major, but also pointed to the conclusion that students' interests in mathematics-related careers were relatively stable over the course of the study. The authors concluded gender differences in mathematics self-efficacy and differing high school mathematics coursework, but not differences in mathematics performance, combined in influencing interest in mathematics-related careers:

> The decision to enter a math/science major was in large part a function of preexisting efficacy and vocational interest patterns. It is apparent that these young women received qualitatively different high school experiences, believed less in their ability to successfully perform math/science tasks, and consequently expressed less vocational interest than young men in mathematics. (Lapan et al., p. 288)

The questions of potential gender or racial differences in mathematics self-efficacy was also considered by Pajares and Kranzler (1995), who found no differences between male and female high school students' mathematics performance, self-efficacy, or general mental ability, but found female students reported higher levels of mathematics anxiety. The authors give evidence to suggest observed levels of mathematics anxiety have only weak direct influences on performance, whereas general mental ability and mathematics self-efficacy display strong and approximately equal effects on both anxiety and mathematics performance for students of both genders.

Similar to the Hackett and Betz (1989) and Pajares and Kranzler (1995) studies, Chen (2003) found no differences in mathematics self-efficacy between boys and girls in middle school, but did find boys expressed more *overconfidence*—tendencies for self-efficacy ratings to exceed performance on matched tasks than girls did in making self-efficacy judgments. Similar overconfidence was noted by Pajares and Kranzler as differing along both gender and race—high school boys were more overconfident than girls, and African American high school students were more overconfident than White high school students. Recently, Chen and Zimmerman (2007) suggested the lower performance of U.S. middle school students when compared to the performance of Taiwanese middle school students may be at least partially explained by a greater tendency for U.S. students (both male and female) to be overconfident in reporting self-efficacy judgments.

Sources of Mathematics Self-Efficacy and Gender

Bandura (1997) proposed that an individual's self-efficacy in a domain such as mathematics develops through experience in the domain and the individual's perceptions of four sources of information: (a) authentic mastery experiences, (b) vicarious experiences, (c) verbal or social persuasions, and (d) emotional and physical states. Both exploratory factor analyses (Lent, Lopez, Brown, & Gore, 1996) and experimental interventions (e.g., Hackett, Betz, O'Halloran, & Romac, 1990) have supported Bandura's theory that self-efficacy beliefs are based on the four sources (see Usher & Pajares, 2008 for a synthesis of the literature). Nonetheless, the relative influences of the four sources on the self-efficacy of individuals may differ substantially across domains, across individuals, and even within an individual when considering self-efficacy for different areas of competence (Zeldin, 2000).

The ways in which students experience and come to internalize information from each of the four sources of self-efficacy in mathematics appears to have measurable influences on their success and persistence in mathematics (Usher & Pajares, 2008). Students' reports of their perceptions of each of the four sources of self-efficacy have been linked to mathematics performance (Lopez & Lent, 1992), interest in mathematics-related careers (Lent et al., 1994), and gender differences in mathematics performance and self-efficacy (Campbell & Hackett, 1986).

Mastery, or performance, experiences are widely considered to be the most influential source of self-efficacy for individuals in most learning domains:

> Authentic mastery of a given task can create a strong sense of efficacy to accomplish similar tasks in the future. Alternatively, repeated failure can lower efficacy perceptions, especially when such failures occur early in the course of events and cannot be attributed to lack of effort or external circumstances. Continued success, on the other hand, can create hardy efficacy beliefs that occasional failures are unlikely to undermine. (Zeldin & Pajares, 2000, p. 216)

Lent and colleagues' (1996) factor analysis of college students' responses to a sources of self-efficacy survey identified mastery experiences as so dominant in self-efficacy formation as to lend some support to utilizing a two-factor model for sources of self-efficacy in statistical modeling: Mastery Experiences and Other. Showing the potential for proximal mastery experiences as having almost immediate effects on self-efficacy, Hackett and colleagues (1990) documented that undergraduate psychology students' success or failure on mathematics tasks directly influenced their self-efficacy ratings on subsequent tasks. However, the authors found no effects of these mastery experiences on students' interests in mathematics-related careers or more general academic self-efficacy, which suggested that experimental manipulation of self-efficacy may have limited lasting effects.

While mastery experiences have received substantial research attention as the primary source of self-efficacy, little research has addressed students' use of the three other sources of self-efficacy, especially learners' perceptions of vicarious experiences and emotional and physical states. Vicarious experiences are thought to influence self-efficacy through observational learning mechanisms: if one observes others succeed or fail after attempting a mathematics task, for example, it may influence his or her self-efficacy to complete a similar task successfully (Bandura, 1997). The construct of vicarious experiences as a source of self-efficacy seems particularly suited for qualitative techniques, but the review of literature identified no qualitative investigations of the role of vicarious experiences as a source of mathematics self-efficacy.

One important qualitative inquiry into the sources of mathematics self-efficacy is Zeldin's (2000) investigation of men and women in mathematics-related careers. Zeldin interviewed 10 men and 15 women in mathematics-intensive technical careers regarding their career choices and early experiences with mathematics. To develop a naturalistic, emergent theory on participants' experiences of the four sources of self-efficacy, Zeldin never specifically asked participants about self-efficacy or the four sources. Zeldin's analysis of the participants' career narratives suggested that men relied primarily on mastery experiences, especially in early college mathematics coursework, in forming the self-efficacy to pursue a mathematics-related career. Women's mathematics career self-efficacy was primarily founded on social persuasions and vicarious experiences. That is, women developed beliefs in their capabilities to become a mathematician, chemist, or computer programmer primarily through encouragement from friends, family, and respected others, as well as by internalizing the belief that other women's success in mathematics meant they could succeed too: "Women rely on relational episodes in their lives to create and buttress the confidence that they can succeed in gender-unfriendly, male-dominated, domains" (Zeldin, 2000, p. 2). This difference may be quantifiable—Lent, Lopez, and Bieschke (1991) found differences in sources of self-efficacy helped explain gender differences in self-efficacy among college students.

A secondary aim of the study was to add to Zeldin's (2000) findings into the sources of self-efficacy in men and women in mathematics-related careers to the specific arena of performance in advanced mathematics. Though women complete approximately equal numbers of advanced courses in high school (Davenport, Davison, Kuang, Ding, Kim, & Kwak, 1998), women undergraduates have been historically underrepresented in advanced college mathematics coursework and have been reported to be historically much less likely than men to express interest in pursuing a graduate mathematics degree (Mura, 1987). In contrast to Zeldin's career-level investigation of the sources of mathematics self-efficacy, this dissertation study investigated the self-efficacy and calibration of men and women enrolled in advanced college mathematics coursework through both quantitative and qualitative methods. Through task-based interviews, the qualitative inquiry also helped to extend Zeldin's findings as well as triangulate the quantitative findings regarding potential gender differences in performance, self-efficacy, and calibration.

Calibration

In educational psychology, a propensity toward overconfidence or underconfidence in one's self-evaluations indicates poor *calibration*, which is defined as the accuracy of evaluative judgments in relation to performance on similar or identical tasks (Schraw, Polenza, & Nebelsick-Gullet, 1993). A person has good calibration in a domain if his or her confidence levels for tasks in the domain align well with subsequent performance; poor calibration means substantive discrepancies between confidence ratings and actual performance (Pajares & Kranzler, 1995). Calibration is a component of *metacognition*—knowledge about, or efforts to regulate and monitor, one's thinking (Schunk, 2004)—in the sense that calibration indicates "how aware individuals are of what they do and do not know" (Stone, 2000, p. 437).

Studies of calibration usually address either *prediction calibration*—the accuracy of self-efficacy judgments made prior to attempting a task (e.g., Chen, 2002)—or *postdiction calibration*, which refers to confidence ratings after completing a task (e.g., Lin & Zabrucky, 1998). Though typically studied using disparate theoretical perspectives (Schraw, 1995), a review of research into both prediction and postdiction calibration uncovered four common themes: (1) adults typically display moderately poor calibration in the form of overconfidence on difficult tasks and underconfidence on easier tasks, (2) calibration is influenced by task difficulty, (3) calibration is conceptually and empirically distinguishable from self-efficacy and outcome expectancies, and (4) calibration is associated with academic performance, especially in mathematics.

Prediction and Postdiction Calibration in Cognitive Science

Calibration research began with experimental studies in the 1950s which demonstrated doctors were consistently overconfident in judging the accuracy of their diagnoses, especially when experimenters provided little feedback on the accuracy of prior diagnoses (Kahneman, Slovic, & Tversky, 1982). Researchers then contrasted the calibration of doctors with the calibration of other professionals that regularly express confidence in statements, such as meteorologists and stock analysts, and later focused on adults' calibration on general knowledge tasks (Lichtenstein, Fischoff, & Phillips, 1982).

The cognitive science approach to calibration research views calibration as a probabilistic "feeling-of-knowing accuracy" (Schraw, 1995, p. 326). Researchers ask participants to rate feelings of confidence for their answers to a series of multiple choice questions (e.g., I am 70% confident the answer I gave was correct). These postdiction confidence ratings are later compared to the percentage of correct answers to produce a *calibration curve* for each participant based on the relative difficulty of the tasks (Lichtenstein et al., 1982). Figure 3 shows a typical postdiction calibration curve, including a tendency toward overconfidence on less-difficult tasks and underconfidence on more difficult tasks. This method for evaluating postdiction calibration requires participants to complete many—sometimes hundreds of—tasks with varying difficulty, and studies using this method often incorporate counter-intuitive statements (i.e., trick-questions) in which adults are typically very poorly calibrated (Stone, 2000).

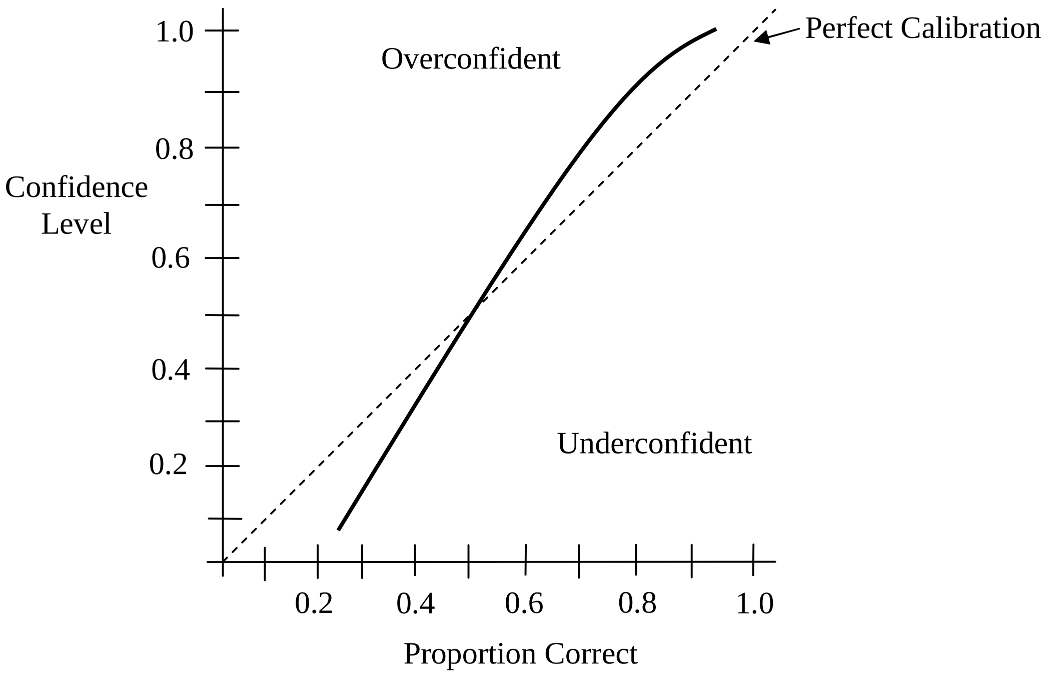

Figure 3. Typical adult postdiction calibration curve for general knowledge tasks (Adapted from Stone, 2000, p. 441).

O'Connor (1989) conducted an extensive review of calibration studies completed under the (probabilistic) cognitive science approach. Linking the results to contingency models in the behaviorist learning paradigm, O'Connor suggests adults' prediction and postdiction calibration is linked to the context of the tasks, the rater's familiarity with the task requirements and topic of interest, and the adequacy of feedback on the results of prior similar tasks. In light of the value cognitive science researchers place on probabilistic alignment between confidence ratings and performance, O'Connor cautions that assigning accurate probability values to feelings of confidence is a skill that few people develop without practice. However probabilistically inaccurate, he notes that confidence ratings from even inexperienced adult participants are typically reliable, with reported test-retest and split-half correlation coefficients in adults' confidence ratings range from r = .72 to r = .85 in experimental calibration studies.

O'Connor's (1989) review identified several exceptions to the often-reported trend toward overconfidence in difficult calibration tasks, including excellent reported calibration curves in contexts where participants had high task familiarity, such as weather forecasting by meteorologists and prediction of course grades by college students. In the specific situation of college mathematics, however, Mura (1987) found students often overestimated their final grades. Interestingly, men overestimated their final grade in college mathematics classes 61% of the time and underestimated only 13% of the time, while women overestimated their grade 51% of the time and underestimated 23% of the time.

The accuracy of an adult reader's beliefs in his or her understanding of textual material, or *metacomprehension accuracy* (Thiede & Anderson, 2003), is a calibration construct that has received substantial attention in reading education research (Zhao & Linderholm, 2008). This form of calibration has been operationalized as alignment between a reader's confidence in their responses to reading comprehension tasks and their performance on the tasks. Maki, Shields, Wheeler, and Zacchilli (2005) found that prediction calibration in reading is strongly correlated to postdiction calibration (r = .83, $p < 0.01$), and that *bias*, or the signed difference in confidence ratings prior to taking a test and subsequent test performance, was the most predictive measure of metacomprehension accuracy. Moreover, task difficulty significantly affected metacomprehension accuracy—although low-ability readers were more overconfident than were high-ability readers, both high- and low-ability readers were more overconfident in their understanding of difficult reading passages than in their understanding of easier passages.

In their review of reading calibration (i.e., metacomprehension accuracy), Zhao and Linderholm (2008) cite evidence that calibration is influenced primarily by readers' familiarity with the content of reading passages and performance expectations, suggesting an *anchor and adjustment* perspective on the self-efficacy judgments of readers who expect to engage in reading comprehension tasks.

> [Readers] may anchor their judgments on pre-formed performance expectations and then adjust their judgments based on experiences with current tasks. Adjustments tend to be insufficient, so the final judgment values are biased toward the anchor. This anchoring and adjustment mechanism can be used to explain how metacomprehension judgments are influenced by both experiential cues and pre-formed performance expectations but seem to be affected by the latter to a greater extent. (Zhao & Linderholm, 2008, p. 7)

Zhao & Linderholm's (2008) anchor and adjustment approach to metacomprehension accuracy assumes adult learners use past reading and performance experiences to form general outcome expectations. Readers then adjust those expectations based on task-specific cues in forming self-efficacy judgments for individual tasks. This perspective situates calibration as a result of self-regulatory cognitive monitoring processes (Thiede & Anderson, 2003) that rely heavily on prior experiences. Although this fits into the Lent, Brown, and Hackett (1994) general model of academic choices discussed earlier (see Figure 2), a social cognitive view of calibration suggests self-efficacy influences anchor and adjustment processes by exerting powerful effects on motivation, effort, and persistence (Bandura, 1997). Moreover, social cognitive theory provides an alternative view of overconfidence in self-efficacy judgments.

Interpretations of Overconfidence

An assumption that overconfidence is maladaptive for performance underscores much of the calibration research. Lichtenstein and Fischhoff (1980), for example, conducted experiments to improve the calibration of adults on general knowledge tasks using only the rationale that externally adjusting an assessor's confidence ratings is very difficult, "so one would like to have probability assessors whose assessments are unbiased to begin with" (p. 150). In taking the perspective of people's use of intuition during situations of uncertainty, Fischbein (1987) sees probabilistic overconfidence as a very general tendency to hold unrealistically high feelings of confidence: "we are inclined to admit, with a feeling of absoluteness, statements which are objectively only weakly supported by empirical data or logical arguments" (p. 29). In Fischbein's view, patterns of probabilistic overconfidence simply reflect internal cognitive tendencies toward feelings of certainty that do not coincide with the probabilistic form of certainty that is so highly-valued by cognitive science researchers.

Social cognitive theorists suggest slight to moderate overconfidence is actually a good thing in many learning situations, because a belief that one is capable of accomplishing a task increases motivation and effort on the task, which in turn expands the possibilities of what someone can actually accomplish (Bandura, 1997). Leading self-efficacy researcher Frank Pajares, in an interview with Madewell & Shaughnessy (2003), cautions against viewing overconfidence as academically maladaptive:

> What seems clear, however, is that we should not tinker with overconfidence. Tailhard de Chardin wrote that "it is our duty as human beings to proceed as though the limits of our capabilities do not exist." Who can ever assess a student's full potential with complete accuracy? Students surprise us all the time, just as we surprise ourselves. We should be careful about attempting to "calibrate" a student's self-efficacy beliefs. Improving students' calibration—the accuracy of their self-efficacy beliefs—is an enterprise fraught with potential dangers. Remember that the stronger the self-efficacy, the more likely are persons to select challenging tasks, persist at them, and perform them successfully. Efforts to lower students' efficacy beliefs should be discouraged. (p. 397)

Prediction Calibration in Social Cognitive Theory

Calibration in social cognitive theory refers to relationships between self-efficacy judgments and performance on a relatively narrow range of tasks in a specific domain (Bandura, 1997). Although the postdiction methodology common to the probability-based cognitive science approach to calibration has been used extensively as an alternative to the social cognitive approach to calibration in social cognitive research (Bouffard-Bouchard, 1990), recent research has emphasized the latter form of prediction calibration (e.g., Chen & Zimmerman, 2007; Pajares & Kranzler, 1995).

Unless otherwise noted, in the remainder of this document *calibration* refers to the social cognitive view of prediction calibration which relates self-efficacy judgments and performance on similar or identical tasks.

The social cognitive theory method for assessing calibration allows for distinctions between *accuracy* and *bias* (Schraw, 1995). Using common scales for performance and self-efficacy (e.g., Chen, 2003 used scores from 0 to 5), an individual's bias on a task is the signed difference of the performance score and self-efficacy rating on the task. That is, for a given task, *bias* = *self-efficacy* – *performance*, so that a positive bias score indicates overconfidence on a task, a bias score of 0 indicates perfect calibration, and negative bias indicates underconfidence. Accuracy is calculated by subtracting the magnitude of bias scores from the maximum possible performance score on an item (Pajares & Graham, 1999): *accuracy* = *maximum performance score* – | *bias* |. Thus, accuracy values fall between 0 and the maximum performance score, with greater values indicating better calibration on an item.

The sensitivity of calibration measures to assessment formats is particularly important because “if improved calibration is in part a function of self-efficacy assessment, then the assessment itself becomes a useful intervention to help students with this metacognitive capability” (Pajares & Miller, 1997, p. 216). Pajares and Miller used a crossed experimental design to assess the mathematics self-efficacy and calibration of 327 middle school students assigned to one of four conditions corresponding to open-ended vs. multiple-choice self-efficacy measures (i.e., continuous vs. ordinal scales) and open-ended vs. multiple-choice mathematics tasks. The authors found no differences in self-efficacy ratings across assessment formats, but found students’ calibration was significantly poorer on open-ended mathematics tasks. Pajares and Miller argue the poorer calibration exhibited by students on open-ended tasks suggests greater validity in calibration assessments based on open-ended performance tasks (i.e., calibration scores can improve by guessing on multiple-choice tasks).

The Calibration of Self-Efficacy Judgments in Mathematics

The review of literature identified a number of studies that incorporated prediction calibration of self-efficacy judgments in mathematics, most of which included students in Grades 5-12. Just two studies of mathematics self-efficacy and calibration among college students were found— Bouffard-Bouchard’s (1990) investigation of mathematics calibration among 64 Canadian college students, and Hackett and Betz’s (1989) study of mathematics self-efficacy, calibration, and college majors among 262 U.S. college students. (The results of these two studies were addressed in the prior section on Self-efficacy in Mathematics Education.)

Some particularly cogent studies of mathematics self-efficacy and calibration at the middle and secondary levels include: (1) Ewers and Wood's (1993) study of gender differences among gifted boys and girls in 5th grade mathematics, (2) Pajares and Kranzler's (1995) study of general mental ability, gender, and ethnicity among 329 students in an urban high school, (3) Pajares & Graham's (1999) investigation of academic motivation, gender, anxiety, and gifted status among 273 first-year middle school students, and (4) Chen's (2002, 2003; Chen & Zimmerman, 2007) studies of prior mathematics achievement, effort judgments, and self-evaluations following mathematics tests among middle school students in the U.S. and Taiwan. In particular, path models tested by Pajares and Kranzler (1995) and Chen (2003) heavily influenced the design of the dissertation study reported in subsequent chapters.

Figure 4 illustrates the hierarchical model used by Chen (2003) in her path analysis of mathematics achievement, self-efficacy, and calibration. Chen found moderate effects of prior achievement on mathematics self-efficacy ($\beta = .42$) and calibration ($\beta = .44$), but even stronger effects of mathematics self-efficacy on performance ($\beta = .50$) and of calibration on mathematics performance ($\beta = -.63$). Inclusion of calibration in a linear regression model greatly improved the model—self-efficacy alone explained 25% of the variation in mathematics performance, while self-efficacy and calibration combined to explain 65% of the variation in mathematics performance. Self-efficacy had a very large direct influence on students' post-test self-evaluations of performance ($\beta = .77$), suggesting that U.S. middle school students' self-beliefs in their mathematics capabilities strongly influence their self-evaluations of performance after completing mathematical tasks.

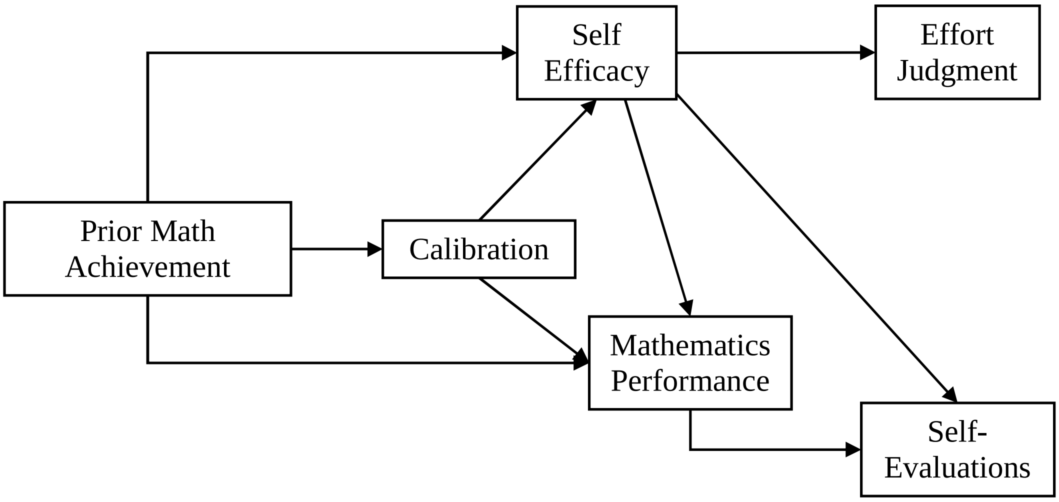

Figure 4. Chen's (2003) path diagram (all paths significant at α =.05). Hypothesized effects of gender on calibration, self-efficacy, and mathematics performance were not supported by the data, so were omitted.

Though Pajares and Kranzler (1995) investigated the calibration of high school students' mathematics self-efficacy judgments, they did not incorporate calibration measures into their path diagram for mathematics performance (see Figure 5). The authors did note, however, that calibration scores were moderately correlated with general mental ability (r = .42) and mathematics performance (r = .67), but were only very weakly correlated with self-efficacy ratings (r = .17). This finding echoes Chen's (2003) finding that there was only a very weak effect of calibration on mathematics self-efficacy (β = -.01), suggesting that mathematics self-efficacy and calibration exhibit independent effects on mathematics performance.

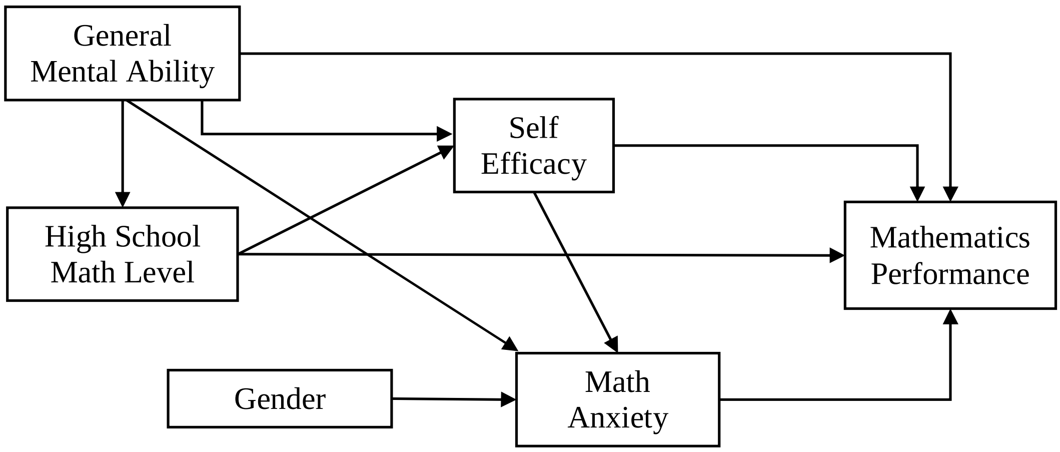

Figure 5. Pajares and Kranzler's (1995) path diagram (all paths significant at α =.01).

It also worth noting that, although Pajares and Kranzler's (1995) model incorporates general mental ability, high school mathematics level, and gender as variables influencing mathematics performance, the 61% of mathematics performance variance explained by their model is similar to the 69% of variance in mathematics performance explained by Chen's (2003) model that included only self-efficacy, calibration, and prior mathematics performance as variables influencing mathematics achievement.

Hypothesized Model for Advanced Mathematics Performance

Informed by the review of literature, the quantitative strand of the study included a hypothesized structural model for mathematics exam performance that incorporates (1) high school mathematics achievement as an exogenous variable, and (2) the extent of mathematics in participants' college major, (3) calibration, and (4) self-efficacy as endogenous variables. Formal hypotheses corresponding to the structured path diagram, shown in Figure 6, appear at the end of this chapter. Similar to Chen (2003), the study also included multivariate analysis of variance (MANOVA) tests for potential differences in endogenous variables by gender and the difficulty of exam items.

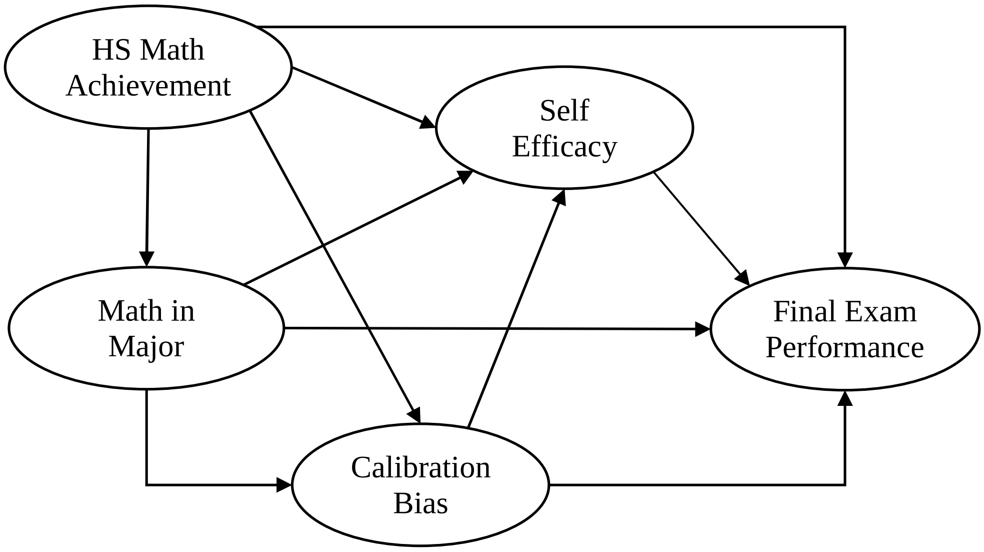

Figure 6. Hypothesized structural path model for performance in advanced mathematics. Arrows indicate unidirectional effects.

Rationale for the Hypothesized Model

The hypothesized model for advanced mathematics performance among college students represents a blending of constructs and directional influences arising from, and supported by, related literature. The model is based primarily on Chen's (2003) path analysis of performance, self-efficacy, calibration, and effort among middle school students, but also includes the amount of mathematics in participants' college major as an endogenous variable. Moreover, the model omits two constructs used in Chen's study—post-test self-evaluations of effort and performance—due to findings from the review of literature that the two measures add little conceptual or predictive value to the model beyond effects of pre-test self-efficacy evaluations and calibration, respectively. The inclusion of the amount of mathematics in participants' college major reflects the literature review findings of marked differences in mathematics-related career choices associated with variation in mathematics self-efficacy. That is, students' choices of college major indicate a broad form of mathematics self-efficacy in the sense that college students are expected to typically choose a major they believe they are capable of completing.

In hopes of placing the study in the context of the related literature, Table 4 summarizes constructs identified appearing in studies that have investigated the relationships between mathematics self-efficacy and academic performance, along with indications of which constructs are addressed by the hypothesized model.

Table 4.

Summary of Constructs in Studies of Mathematics Self-efficacy and Performance

Construct	Example References	Included
Self-Efficacy	Chen, 2003; Hackett & Betz, 1989	X
Self-Concept	Pajares & Miller, 1994	
Calibration	Chen & Zimmerman, 2007; Bouffard-Bouchard, 1990	X
Math Anxiety	Pajares & Graham, 1999	
Effort	Chen, 2003	
Persistence	Pajares & Graham, 1999	
Posttest Self-Evaluations	Chen, 2003; Chen, 2002	
Task Difficulty	Chen, 2003; Maki et al., 2005	X
Assessment Format	Pajares & Miller, 1997	
Goal Orientations	Elliot & Moller, 2003; Midgley & Urdan, 2001	
Math-Related Career Interests	Lapan, et al., 1996	
College Major	Hackett & Betz, 1989	X
Prior Math Achievement	O’Brien, et al., 1999	X
General Mental Ability	Pajares & Kranzler, 1995	
Gifted Status	Ewers & Wood, 1993; Pajares & Graham, 1999	
Gender	Pajares & Kranzler, 1995; Campbell & Beaudry, 1998	X
Socio-Economic Status	O’Brien, et al., 1999	
Ethnicity	O’Brien, et al., 1999; Pajares & Kranzler, 1995	

The purpose of the preceding review of literature was to summarize self-efficacy and calibration research with the ultimate goal of developing a social cognitive model for advanced mathematics performance among college students in the target population. The hypothesized model was used in both the quantitative and qualitative strands of the inquiry. Nonetheless, review of literature in qualitative inquiries is often both emergent and cyclical (Patton, 2002), and additional review of literature supporting the qualitative inquiry emerged during data collection and interpretation. Partly because nearly all self-efficacy research has been conducted within the quantitative educational research paradigm (Bandura, 1997), the qualitative inquiry was exploratory in nature and was informed by both the quantitative findings and a diverse collection of related literature.

Summary of Literature Review

The preceding literature review outlines a conceptual framework supported by extensive empirical evidence suggesting self-efficacy and calibration have important influences on academic motivation and performance, especially in mathematics. Initial sections of the literature review described the concept of self-efficacy and distinguished it from related motivation constructs such as self-esteem, self-concept, outcome expectancies, and goal orientations. Extensive research in a variety of educational fields has documented self-efficacy as an important predictor of performance, effort, and persistence in mathematics. Following a discussion of self-efficacy research and concepts related concepts, the review summarized research into the concept of calibration, which suggested harboring accurate feelings of confidence is an important aspect of metacognition.

Following synthesis of theoretical and empirical models of mathematics performance from the social cognitive perspective, the review of literature ended with a hypothesized model for performance in advanced college mathematics. Though the review of literature in the qualitative strand of the inquiry was emergent and cyclical, the summary of the research into self-efficacy and calibration helped to lay the foundation for task-based interviews, including rationale for exploring the ways in which sources of self-efficacy and hypothesized relationships between self-efficacy, calibration, and mathematics performance may be characterized among prospective secondary mathematics teachers enrolled in advanced mathematics courses.

Research Questions

Recall the guiding research question: How do self-efficacy and calibration influence the exam performance of students enrolled in the advanced mathematics courses of a secondary teacher preparation program at a mid-sized liberal arts university?

Four quantitative research questions arose directly from the hypothesized model (see Figure 6) for mathematics performance, including one question for each of the endogenous constructs in the model. Moreover, two research questions addressed potential differences in the endogenous variables by intervening variables identified in the review of literature as being especially pertinent to the target population: students' gender and the difficulty of exam items. Finally, a single qualitative research question called for a holistic comparison of quantitative effects to the processes supporting relationships between self-efficacy, calibration, and mathematics performance within the important subpopulation of prospective secondary mathematics teachers.

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects on participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

Q7 In what ways do prospective secondary mathematics teachers' mathematical problem-solving compare and contrast with the hypothesized relationships between self-efficacy, calibration, and performance in advanced mathematics?

Based on the pilot study results and findings from the review of literature, the quantitative hypotheses pertaining to the six quantitative questions included:

H1 High school mathematics achievement will have a moderate positive effect on the amount of mathematics in participants' college major.

H2 Both high school mathematics achievement and the amount of mathematics in participants' college major have small positive effects on participants' calibration.

H3 High school mathematics achievement and the amount of mathematics in participants' college major will have moderate positive effects on self-efficacy. Calibration will have a small negative effect on self-efficacy.

H4 High school mathematics achievement and the amount of mathematics in participants' college major will have small positive effects on mathematics performance. Calibration will have a large negative effect on mathematics performance. Self-efficacy will have a large positive effect on mathematics performance.

H5 There will be no significant difference in self-efficacy, performance, or calibration by gender. There will be significant differences in the amount of mathematics in participants’ college major by gender, with males on average choosing college majors with more required mathematics courses.

H6 There will be no significant difference in self-efficacy by item difficulty. There will be a significant difference in calibration by item difficulty, with a tendency toward overconfidence on more difficult exam items.

The qualitative research question, regarding relationships between self-efficacy, calibration, and performance in the problem solving of prospective secondary mathematics teachers, was addressed using task-based interview methods and interpreted in the context of the conceptual framework derived from the review of literature. This included an emergent, naturalistic inquiry design which aimed to preclude *a priori* hypotheses (Patton, 2002) of potential themes that would emerge from the data. The aim of the qualitative research question was to help clarify the broad statistical trends identified in the quantitative research questions in the subset of participants with intentions of becoming secondary mathematics teachers.

CHAPTER III

METHODOLOGY

The purpose of the preceding chapters was to build the rationale and a conceptual foundation for a study of self-efficacy, calibration, and exam performance among college students enrolled in the mathematics courses required by a secondary mathematics teacher preparation program. The research problem, research questions, and significance of the study helped established a need for the study. Then, a basis for the study was established through an extensive review of self-efficacy and calibration literature and the development of a social cognitive model for performance in advanced mathematics. This suggested hypotheses regarding expected statistical effects among self-efficacy, calibration, high school mathematics performance, gender, the amount of mathematics in students' college major, the difficulty of test items, and performance on advanced mathematics exams.

Research Questions and Model

The focus of the research design was a single guiding research question: How do self-efficacy and calibration influence the exam performance of students enrolled in the advanced mathematics courses of a secondary teacher preparation program at a mid-sized liberal arts university? To further narrow the scope of the investigation, seven research questions (Q1-Q7) accompanied a hypothesized structural path model (Figure 7) in guiding the research design and methodology.

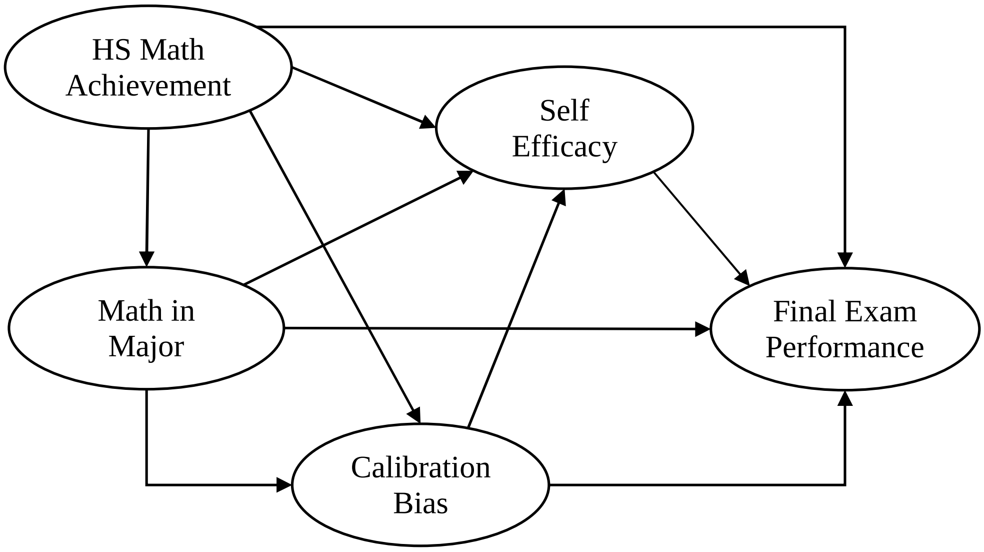

Figure 7. Hypothesized structural path model for performance in advanced mathematics. Arrows indicate unidirectional effects.

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects on participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

Q7 In what ways do prospective secondary mathematics teachers' mathematical problem-solving compare and contrast with the hypothesized relationships between self-efficacy, calibration, and performance in advanced mathematics?

To acknowledge the ways in which the ontological and epistemological orientations of the researcher influenced the research design and methodology, the following section describes the researcher stance. This is followed by a summary of the theoretical perspective, the setting, data collection and analysis procedures, and efforts to gather evidence in support of reliability, validity and trustworthiness for the study.

Researcher Stance

One characteristic that distinguished this study from other investigations of mathematics self-efficacy was the ontological and epistemological orientation informing the research design. Academic self-efficacy research has been conducted almost exclusively within the quantitative research paradigm (Lightsey, 1999), and the review of literature is dominated by discussion of cross-sectional and quasi-experimental studies of psychological constructs and mathematics achievement. As Simon (1999) explains, this reflects the post-positivist, neo-behaviorist history of social cognitive theory, but the theory has increasingly shifted to a social-constructivist orientation toward knowledge construction which includes concern for the many nuances of co-constructions of self-efficacy, such as cultural efficacy and the influences of social norms and valued practices (Bandura, 1997). Thus, the research design incorporates many of the methods and constructs from social cognitive theory while retaining sensitivity toward the provisional nature of mathematics education research findings. The researcher ascribes to a pluralistic ontological orientation (Schwandt, 2001), which means there may be multiple “true” interpretations of human behavior depending on the context and viewpoint of those who might observe such behavior.

The researcher stance was also informed by a pragmatic orientation (Patton, 2002) toward educational knowledge claims in the sense that claims about learning and academic motivation were considered useful by the ways in which they contribute to practical understandings of teaching, learning, policy, and research. As in this study, pragmatic epistemologies are often evident in mixed methods research designs (Creswell, 2003). The research questions drove the choice of methods, and quantitative and qualitative viewpoints were seen as complementary and potentially equally powerful in helping to answer the research questions. In particular, it is important to stress that the relatively larger quantitative component in the study, and commensurate choices in data analysis and voice, were not intended to indicate that the researcher places greater value in findings derived from statistics than findings derived from qualitative methods.

The narrative voice in this study follows conventions in quantitative research that include omission of personal pronouns (e.g., “I”, “me”, “we”). This is intended to maintain consistency throughout the manuscript and was not intended to indicate a post-positivist research orientation. Moreover, the researcher’s involvement with participants is probably best characterized as close to the “observer” dimension of the participant-observer continuum in qualitative research (Creswell, 2007). This means a relative lack of engagement in the students’ mathematics experiences, which in turn limits the potential for in-depth, holistic, accounts of students’ development of course-specific self-efficacy and calibration. Nonetheless, the researcher is sensitive to the calls for increased reflexivity (Glesne, 2006) in qualitative research and engaged in the research with intentions to make researcher biases explicit in the discussion of findings.

Theoretical Perspective

Conceptual Framework

The terms, concepts, and psychological constructs used in this study are primarily based in Bandura's (1997) theory of self-efficacy and social cognitive learning theory. In addition, the cognitive science research into prediction and postdiction calibration, though approached through a different learning theory, informs much of the research. For example, the distinction between calibration bias and calibration accuracy was developed by Schraw (1995), whose cognitive information processing conceptual framework differs substantially from that of social cognitive theorists. Prominent concepts used in this study include self-efficacy, calibration, advanced mathematics, sources of self-efficacy, college major, exam performance, high school mathematics achievement, test item difficulty, gender, and prospective secondary mathematics teachers. Each of these concepts was described in detail in the introduction and review of literature, but the operational definitions are yet to be explicated.

Definitions of Constructs and Indicators

The purpose of this section is to describe operational definitions for the constructs and indicator variables used in the quantitative strand of the investigation. Some of the constructs, such as self-efficacy and high school mathematics achievement, have alternative conceptions in educational research, so the following definitions were regarded as local definitions of the constructs for data collection and analysis purposes and were not intended to encompass the full range of potential meanings for the terms. See the review of literature in Chapter II for additional detail on the diverse conceptions of the constructs.

High School Math Achievement is a latent construct indicated by three measures of students' performance and course taking prior to attending college. Since college readiness is one goal of high school education in the U.S., students' score on the mathematics portion of the ACT college readiness exam was one indicator of high school mathematics achievement and was denoted *ACT Math*. If only an SAT score was available, the score was converted to its approximate ACT equivalent (College Board, 2008). ACT scores were gleaned from institutional records and could range from 11 to 36. High school grade point average, denoted *HS GPA*, also provided a continuous indicator of high school achievement, with a theoretical range of 0 to 4. Finally, *HS Self* referred to students' self-reported assessment of their performance in high school mathematics courses. Students' responses to the question "Which of the following best describes how well you did in your high school math courses?" were coded on a Likert-type scale ranging from 1 = really bad to 7 = excellent.

Self-efficacy is a latent construct associated with students' confidence in their abilities to correctly complete examination items in the minutes just prior to taking a regular exam. Indicators of this construct include numeric records of students' responses on seven pre-exam survey items, each recorded in the interval 0 to 5.

Since the instruments used to attain indications of self-efficacy were different for each final exam and each course, indicators of self-efficacy were constructed by ranking survey items by ascending class means. That is, the item on each of surveys that resulted in the lowest mean self-efficacy rating among students in a given section corresponded to the indicator variable SE Level 1. Students' self-efficacy ratings on the item for which the mean self-efficacy in the class was next highest formed the indicator label SE Level

2, and so on. This assignment of students' responses into indicators based on ascending class means formed seven indicators of exam self-efficacy ranging from a student's self-efficacy on the item for which the class was least confident (SE Level 1) to the student's self-efficacy on the item for which the class was most confident (SE Level 7).

Math in Major means the amount of required mathematics content in the students' chosen college major. The total number of required semester credits with a university catalog prefix of MATH in a student's college major (range = 3 to 45), labeled *Required Math,* represented the sole indicator of the Math in Major construct.

Final Exam Performance, or simply *Performance,* is the latent construct associated by a student's achievement on a regular in-class final exam. Performance on individual exam items was scored on a dichotomous scale (0 = incorrect, 5 = correct).

As in the indicators of Self Efficacy, Level indicators of performance were formed by ranking seven final exam items according to ascending within-class mean performance during the final exam. For a given final exam, seven items were randomly sampled to be representative of the difficulty of items *not* included on the self-efficacy survey. That is, the mean class performance on non-self-efficacy items was calculated, items were stratified into seven quantile groups based on the rank-ordering of items, and a single item was sampled from within each of the seven quantile groups. For example, *Performance Level 1,* referred to students performance on the sampled final exam item with the lowest mean within-class performance. That is, Performance Level 1 represented students' performance on the "hardest" sampled final exam item, while Performance Level 7 represented students' performance on the "easiest" sampled final exam item.

Calibration is the latent construct indicated by the difference in students' self-efficacy rating and performance for the seven tasks on the final examination. Calibration bias scores for an individual task could range from -5 (underconfidence) to +5 (overconfidence) for each task. As in the operationalization of self efficacy and performance indicators, level indicators of calibration bias, labeled *Bias Level 1* through *Bias Level 7*, were constructed by including students' bias scores on exam items corresponding to ascending within-class mean calibration bias scores.

Gender is the self-reported sex of participants and was nominally coded 1 = Female and 2 = Male.

Item Difficulty refers to the mean class performance of students on the final exam items presented to students on the self-efficacy surveys. Similar to the procedure used to order indicators of Final Exam Performance, the item difficulty for tasks presented on self-efficacy surveys were sorted rankings of within-class mean performance scores. For ease of interpretation, however, the rankings were reverse-ordered to represent descending mean performance. For example, a survey item with Difficulty Level 1 was the "easiest" survey item in the sense that highest percentage of people correctly completed the item. Difficulty Level 7, in contrast, would be considered the "hardest" survey item.

Structural Equation Modeling

The study incorporated a path model (Figure 7) for mathematics achievement that includes a saturated path diagram which posits multiple directional effects among several latent *constructs*, or unobservable variables. Constructs that are endogenous to (predicted by) one construct are often exogenous to (predictive of) another construct, and measurement of these latent constructs necessarily permits the likelihood of measurement error. This measurement error and nesting of multiple dependent variables in directional relationships violates assumptions of standard multiple linear regression techniques (Snedecor & Cochran, 1989) and suggests structural equation modeling techniques. Structural equation modeling allows for simultaneous estimation of directional effects and measurement error among observed variables, called *indicators*, and latent constructs using through a blending of regression and common factor analysis of observed correlation structures (Schrieber, 2008). Structural equation modeling is thus a technique to analyze multiple directional effects among several latent variables, or constructs, in cases where each such construct can be approximated through one or more ordered indicator variables (Loehlin, 1987).

In structural equation modeling, there are three important types of diagrams used to explain the constructs and indicators in the model. A *structural path model* encapsulates the hypothesized "paths" or directional effects between latent variables, and must be supported by theory and prior research (Hair, Anderson, Tatham, & Black, 1998). The path model for this study is shown in Figure 7 and is supported by the review of literature in Chapter II. The second kind of diagram is called a *structural measurement model* (Byrne, 1998) and includes specification of the latent constructs which serve as common factors influencing the observed indicator variables in the model. While there is no fixed requirement for the number of indicator variables that "load onto" a construct in the measurement model, Hair and colleagues (1998) suggest validity of structural equation modeling is typically best when most constructs have 3 to 7 indicators. Finally, the *structural model* diagram specifies all the hypothesized relationships by including both the effects between latent constructs (path model) and the effects of latent constructs on indicator variables (measurement model).

A diagram of the hypothesized structural model in this study is shown in Figure 8. For convenience, measurement errors are sometimes omitted from drawings of the structural model diagram, and the structural path model is sometimes referred to simply as the structural model. The convention of denoting latent variables as ovals and indicator variables as rectangles (Schrieber, 2008) is retained throughout the report.

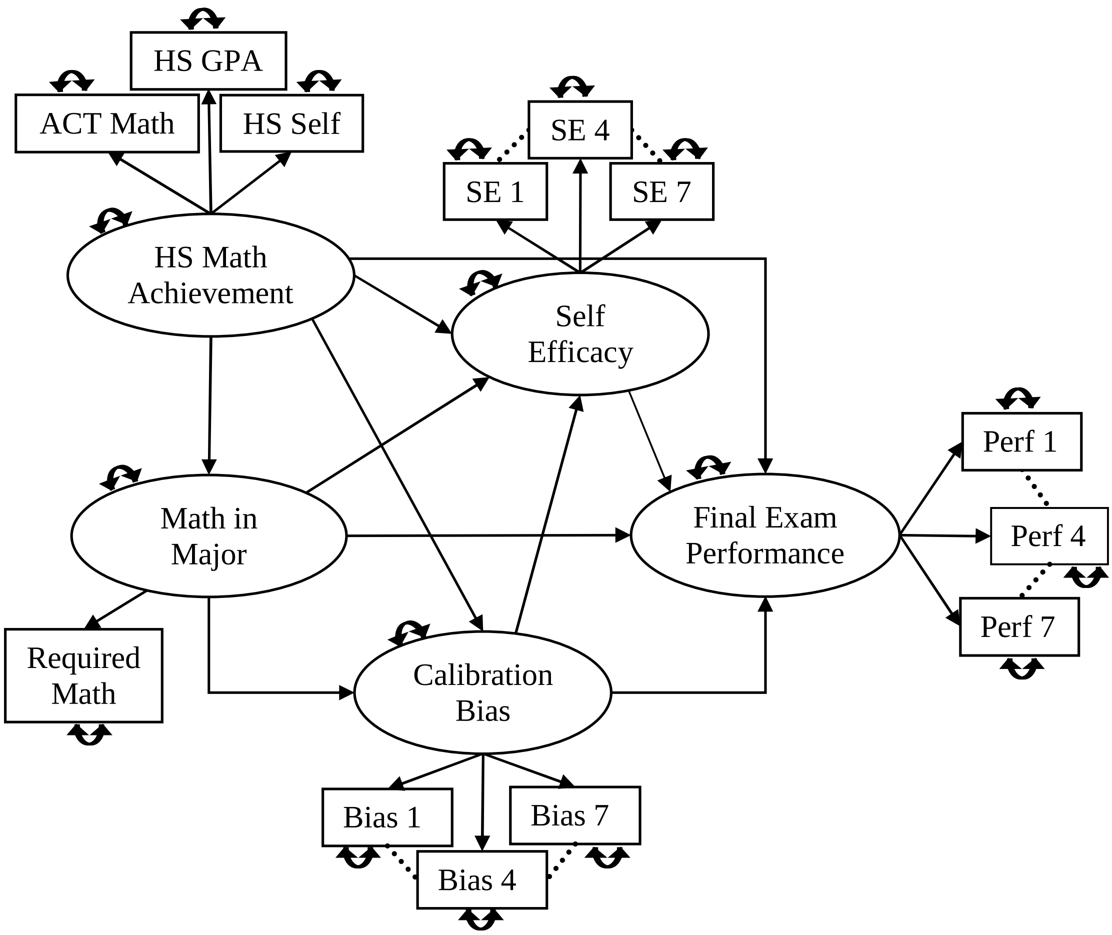

Figure 8. Diagram of hypothesized structural model. Arrows between latent constructs – drawn as ovals – form the path model. Arrows from latent constructs to (observed) indicator variables – drawn as rectangles – form the measurement model. Measurement errors are indicated as small bidirectional arrows.

Research Setting

The potential significance of this study was partly derived from the atypical mixture of students' college majors at the research site. In relation to national norms, the mathematics department at the research site serves comparatively large numbers of preservice teachers, and many of the mathematics majors have chosen a secondary mathematics education. Before going into details regarding the population of students enrolled in advanced mathematics courses at the research site, it is informative to consider the national context concerning enrollment in advanced mathematics courses.

The National Context

According to the American Mathematics Society's 2005 survey of mathematics departments, advanced mathematics courses such as calculus, differential equations, and linear algebra accounted for 43% (699,000) of the more than 1.6 million total student enrollments in college mathematics courses (Lutzer, Rodi, Kirkman, & Maxwell, 2007). However, the vast majority of this national enrollment comes from students majoring in engineering, computer science, and the physical sciences. Lutzer (2002), for example, found that only a tiny proportion (0.6%) of U.S. incoming college freshman plan to major in mathematics, while about one-fourth (25-30%) of freshmen intend to major in a science or engineering field. Interestingly, the proportion of mathematics majors actually increases from freshman to senior student-populations, with 1% (12,363 of 1,199,579) of all U.S. bachelor's degrees going to mathematics majors in 1998.

Even among the relatively few students majoring in mathematics, there are considerable differences in students' interests and purposes in taking advanced mathematics courses. For example, mathematics departments report many more students majoring in applied mathematics and liberal arts mathematics in U.S. universities than students majoring in mathematics education (Lutzer, Rodi, Kirkman, & Maxwell, 2007). Lutzer and colleagues found 14,610 U.S. college students had declared applied or liberal arts mathematics as their major in 2005, compared to just 3,369 students majoring in mathematics education. Moreover, about 40% of U.S. mathematics majors are female compared to 60% of mathematics education majors (Lutzer et al.). The resulting diverse composition of interests and purposes in advanced mathematics courses poses a challenge to mathematics instructors and has the potential of affecting preservice secondary teachers' performance through mediating motivational factors.

Advanced Mathematics in the Research Site

The secondary mathematics teacher preparation program in the research site requires 12 mathematics content courses including Calculus I-III, Linear Algebra, Discrete Mathematics, Abstract Algebra I & II, Modern Geometry I & II, Mathematical Modeling, Elementary Probability Theory, and History of Mathematics. While the required mathematics content courses reflect traditional content in the preparation of secondary mathematics teachers, there are some atypical characteristics of student enrollment and instructional strategies in the mathematics content courses.

Partly due to the university's liberal arts composition of student majors and special focus on preparing school teachers, a majority of students enrolled in the mathematics content core classes intended to major in mathematics or a related teaching field. Institutional records from the spring semesters of 2007 and 2008 indicated that approximately 42% of all students enrolled in mathematics content courses had declared a major in mathematics, a rather large percentage in light of the previously mentioned national surveys indicating students majoring in mathematics constitute a small minority of advanced mathematics courses (Lutzer et al., 2007). Table 5 summarizes average spring enrollment by major in selected mathematics courses at the local university.

Table 5.

Enrollment by Major in Selected Mathematics Courses at the Research Site

	Calculus I	Calculus II	Elementary Probability Theory	Discrete Mathematics	Other Courses	Total
Mathematics	15.0	19.0	26.5	24.5	103.5	188.5
Elem. Teaching	8.5	3.0	0.0	17.5	19.0	48.0
Biology	28.5	1.0	0.0	0.0	1.0	30.5
Chemistry	13.5	7.5	0.0	1.0	6.5	28.5
Physics	5.0	8.5	1.0	1.5	11.0	27.0
Earth Sciences	7.5	7.0	2.0	0.0	5.5	22.0
Undeclared	10.0	3.5	0.5	1.5	1.5	17.0
Pre-Profess.	8.0	3.5	0.0	0.5	1.0	13.0
Business	6.5	1.0	0.5	0.5	3.0	11.5
All Others	35.0	9.5	1.5	4.5	13.0	63.5
Total	137.5	63.5	32.0	51.5	165.0	449.5

Note. Enrollment counts are averages from the spring semesters of 2007 and 2008.

In addition to the composition of students' majors in advanced mathematics courses, the distribution of emphases for students who have declared a major in mathematics was also atypical at the research site. As of Fall 2008, there were 180 students at the research site who had declared a major in mathematics (see Table 6). Of those, 125 (69%) declared an emphasis in secondary mathematics education, and 99 were female (55%). Moreover, female secondary mathematics majors outnumbered male secondary mathematics majors almost 2 to 1 (82 to 43), indicating that gender differences in self-efficacy, calibration, or performance might relate to potential differences in students' choice of college major emphasis at this university

Table 6.

Distribution of Emphases among Mathematics Majors by Gender, Fall 2008

Emphasis	Male	Female	Total
Applied Mathematics	18	4	22
Liberal Arts Mathematics	20	13	33
Secondary Mathematics Education	43	82	125
Total	81	99	180

The unusual composition of students' majors in advanced mathematics courses at the research site afforded a unique opportunity to investigate (1) relationships between the extent of mathematics in students' choice of college major and their subsequent self-efficacy, calibration, and performance in advanced mathematics coursework and (2) hypothesized roles of self-efficacy and calibration in the mathematics performance of prospective secondary mathematics teachers. Consequently, the task-based qualitative interview protocol was specifically designed to address calibration, self-efficacy, and problem-solving performance within the subpopulation of prospective secondary mathematics teachers.

Research Sample

Sampling Procedures

With consent of instructors, all students enrolled in mathematics courses required for the secondary mathematics education major at the research site were invited to participate in the study; exceptions included only the four sections in which instructors did not administer an in-class final exam. In most cases, students were invited to participate in the quantitative strand of the investigation during a brief visit by the researcher to their classroom between the 6th and 10th week of the semester with informed consent (Appendix A) procedures approved by the Institutional Review Board at the research site. Follow-up letters inviting students to participate were sent by the researcher to students who were not present at the time of the in-class visit and included informed consent documentation.

Expectations of participating instructors included: (1) reviewing self-efficacy surveys tailored to final exam items for representativeness and face validity, (2) providing exams to the researcher several days before administration, (3) allowing consenting students' work on exams to be photocopied prior to grading, (4) working collaboratively with the researcher to construct tasks-based interview prompts for students in their mathematics classes.

The qualitative sampling procedure was a form of criterion-based stratified purposive sampling (Mertens, 2005), with the goal of providing maximum variation in participants' self-efficacy and calibration. Consenting students enrolled in seven sections, encompassing Calculus I, Calculus II, and Elementary Probability Theory, and were asked to complete self-efficacy instruments during their midterm examination in or around the 8th week of the academic semester. Following analysis of the self-efficacy and performance data, those participants who reported a major in secondary mathematics education were ranked based on composite measures of self-efficacy and calibration and ultimately sorted into four efficacy-by-calibration groups: High Self-Efficacy/Good Calibration, High Self-Efficacy/Poor Calibration, Low Self-Efficacy/Good Calibration, Low Self-Efficacy/Poor Calibration. Up to four students from within each of these criterion-based groups were purposely sampled in consultation with the participating instructors with the goal of seeking maximum variation (Patton, 2002).

Frankfort-Nachmias and Nachmias (2000) suggest that criterion-based sampling can introduce a regression effect that may threaten the internal validity of findings because of potential erroneous classifications of participants based on the initial criterion. However, the purposive nature of the qualitative sampling technique, together with the consultations with instructors, was designed to mitigate this threat.

Sample Size

There were 309 students enrolled in the 12 participating sections of advanced mathematics courses (M = 25.8, SD = 6.9). Of the enrolled students, 17 (6%) did not take a final exam and 40 (17%) were enrolled in two or more of the classes, yielding a potential sample of 252 unique students who finished the classes. Of these, 210 (83%) consented to participate; complete final exam and self-efficacy data were available for 195 students. This sample size means that the analysis included data from 77% (195/252) of the students who completed at least one of the 12 participating mathematics classes.

Most (36 of 40) students who were enrolled in more than one participating section were enrolled in two sections, and 4 students were enrolled in three sections. Students enrolled in more than one section were invited to complete self-efficacy surveys and final exams in each of their classes, but only the data from the highest-numbered class in which they were enrolled were included in the analysis.

Table 7 summarizes the distribution of study participants by the class in which the students' self-efficacy and final exam performance were included in the study. Approximate course numbers are also included in Table 7 as indicators of the academic level associated with participating sections. That is, 100-level courses are typically taken by Freshman, 200-level courses are typically taken by Sophomores, and so on. While the study includes data from students completing seven different course titles, about half (49%) of the data comes from students' performance in Calculus I or II.

Table 7.

Distribution of Study Participants by Class

Section	Instructor	Course Title	Course No.	*n*	Subtotal %
1	A	Calculus I	130	20	28
2	B	Calculus I		16	
3	B	Calculus I		18	
4	A	Calculus II	140	18	21
5	A	Calculus II		23	
6	C	Linear Algebra	220	8	4
7	D	Discrete Math	230	21	16
8	D	Discrete Math		10	
9	E	Calculus III	240	11	11
10	E	Calculus III		11	
11	F	Abstract Algebra II	320	22	11
12	G	Probability	360	17	9
			Total	195	100

Instructors

As shown in Table 8, the 12 participating mathematics classes were taught by a total of seven instructors. One instructor taught three of the 12 sections, three instructors taught two sections each, and the remaining three instructors each taught a single section. Five of the seven instructors were tenure or tenure-track mathematics professors, and the remaining two instructors were long-time lecturers at the research site; none of the sections were taught by graduate students, adjunct faculty, or part-time instructors. The instructors averaged 19.0 years of college mathematics teaching experience (*SD* = 11.5, range = 4 to 35).

Participants

Enrollment data available through the research site included several variables which were used to describe the study participants. These included age, academic level, gender, ethnicity, and participants' declared college majors. The paragraphs that follow summarize these characteristics in the context of the undergraduate student population at the research site.

Age

Participants ranged in age from 18 to 49 (*M* = 21.2, *SD* = 4.2). Most of the students (81%) were 18-22 years old, some of the students (11%) were 23-25 years old, or over 25 years old (7%). The fact that study participants were primarily traditionally-aged undergraduate students was reflective of the undergraduate population at the research site, where enrollment records indicate over 90% of new students to the university are under 25 years old (L. Sappington, personal communication, 2009).

Academic Level

The percentages of study participants classified as Freshman, Sophomore, Junior, and Senior were 28%, 28%, 25%, and 18%, respectively. This distribution differs significantly from the proportions of Freshman, Sophomore, Junior, and Senior levels at the research site (χ^2 (3, N = 195) =16.43, $p < .001$), which were 24%, 21%, 23%, and 32%, respectively. In particular, this suggests a slight to moderate under-representation of Seniors in the sample, possibly due to the facts that all of the participating classes were numbered 300 or below and many Seniors are involved in student teaching during the spring semester.

Gender

Study participants were almost exactly equally-distributed by gender (97 female, 98 male). While the observed proportion (50%) of female students enrolled in advanced mathematics courses is substantially higher than national averages (Lutzer et al., 2007), the proportion of female students in the sample was less than the overall proportion (60%) of female undergraduate students at the research site (χ^2 (1, N = 195) = 9.13, $p < .01$).

Ethnicity

As summarized in Table 8, most of the participants self-identified as Caucasian (83%), while some students self-identified as, in order of prevalence, Asian American, Hispanic American, Native American, or African American. The distribution of ethnicities among study participants was not significantly different from the distribution of ethnicities of undergraduates at the research site, χ^2 (5, N = 195) = 2.68, $p = .75$.

Table 8.

Ethnicity of Study Participants and Students at the Research Site

Ethnicity	Participants % (N = 195)	Research Site % (N = 12,475)
Caucasian	83	78
Asian American	5	3
Hispanic American	5	7
Native American	1	1
African American	1	3
Other/Did Not Report	5	8

College Major

Though, all study participants were enrolled in at least one mathematics courses required for the secondary mathematics teacher preparation program at the research site, not all students were mathematics majors. Participants had declared a variety of college majors, most of which were related to sciences, teaching, or both. Table 9 summarizes the distribution of declared college majors among study participants. Approximately half (49%) of study participants declared their primary major in mathematics or mathematics education, including 12% of all students indicating a major in Elementary Education with a concentration in Mathematics and 37% indicating a major in Mathematics. About79% (34/43) of the female mathematics majors chose the secondary teaching concentration, while just 37% (19/30) of the male mathematics majors chose the secondary teaching concentration. Other common majors included Chemistry, Earth Sciences, Physics, Biology, Pre-Program (e.g., pre-medicine, pre-dentistry), and Undeclared.

Table 9.

Declared Primary College Majors of Study Participants (N =195)

Category	Major	Frequency %	Subtotal %
Math Education	Elementary Education – Mathematics	12	35
	Mathematics – Secondary Teaching	23	
Mathematics	Mathematics – Liberal Arts	6	14
	Mathematics – Applied Mathematics	4	
	Mathematics – Statistics	4	
Science	Chemistry	10	28
	Earth Sciences	7	
	Physics	6	
	Biology	5	
Other	Pre-Program	5	23
	Undeclared	5	
	All Others	13	

Interview Participants

As outlined in the procedures section, qualitative interview participants were purposely sampled based on the students' ranking on calibration and performance measures for midterm examinations in seven of the participating classes, including sections of Calculus I, Calculus II, Calculus III, and Probability. Of the 117 consenting students who completed the selected midterm exams and self-efficacy surveys, 22 students had declared a secondary mathematics teaching major. Twelve of these prospective students scored above the median in their class on the midterm exam (High Performance). Similarly, 12 of the secondary mathematics majors scored above their respective within-class medians on the calibration bias measures (High Calibration Bias).

Using the self-efficacy and calibration classifications, a stratified purposeful sample of 12 students was selected from the 2 × 2 array of the Low and High levels of

Performance and Calibration Bias. From this initial sample, two students declined to participate in an interview, leaving 10 interview participants. The task-based interviews took place in the latter half of the semester (Weeks 12-13 of a 16-week semester), lasted between 29 and 65 minutes (*M* = 46.7, *SD* = 10.3), and produced data in the form of students' work on interview tasks and transcribed audio-recordings.

As shown in Table 10, the interview participants were spread approximately equally across Calculus I, Calculus II, and Probability classes, with 3, 3, and 4 students enrolled in the respective courses. Most (8 of 10) interview participants were female. Four interview participants were classified as High Performance and Low Calibration Bias, five were classified as Low Performance and High Calibration Bias, and one was classified as Low Performance and Low Calibration Bias. At the time of the interviews, five of the participants had attained the academic level of Sophomore, four were Juniors, and one was a Freshman. All interview participants were between the ages of 19 and 23.

Table 10.

Task-Based Interview Participants by Course and Other Selected Variables

Participant[a]	Course	Age	Gender	Level	Performance	Calibration Bias
Heather	Calculus I	20	F	Junior	High	Low
Matthew	Calculus I	23	M	Sophomore	High	Low
Megan	Calculus I	19	F	Freshman	Low	High
Justin	Calculus II	23	M	Junior	High	Low
Jackie	Calculus II	20	F	Sophomore	High	Low
Nicole	Calculus II	20	F	Sophomore	Low	High
Sarah	Probability	20	F	Sophomore	Low	High
Jennifer	Probability	21	F	Junior	Low	Low
Emily	Probability	19	F	Sophomore	Low	High
Elizabeth	Probability	21	F	Junior	Low	High

Note. [a]Names are pseudonyms.

Overview of Research Design

Quantitative Strand

The quantitative strand of the study was typical of a structural equation modeling study in education, because it blends simultaneous solutions to multiple linear regression models with analysis of covariance across cross-sectional measures of self-beliefs (Hair et al., 1998). Data collection procedures included a background survey (Appendix B), self-efficacy surveys (Appendix C) in the few minutes just before in-class exams, and photocopies of students’ work on final exam tasks. Self-efficacy scales and calibration bias scores followed procedures that have been incorporated in several mathematics self-efficacy studies (e.g., Chen, 2003, Pajares & Miller, 1994).

One important early question regarding data collection was whether to use a common self-efficacy and mathematics performance measure across all sections of participants in the study. Such a control measure would eliminate variation in item difficulty and academic content due to different examinations in the various courses. However, self-efficacy theorists stress the importance of domain and context specificity when asking learners to assess their mathematical capabilities (Bandura, 1997; Pajares & Miller, 1994). That is, a students' self-efficacy in linear algebra is best measured by asking the student to rate their confidence to complete specific tasks related to their current linear algebra course. Moreover, evidence that self-efficacy ratings may be more reliable when students expect to complete tasks as part of educational requirements (Chen & Zimmerman, 2007) supports the value of providing authentic mastery experiences, such as regular in-class exams, as part of data collection procedures. Thus, self-efficacy surveys and mathematical performance tasks were selected from among the tasks chosen by instructors for in-class exams.

Qualitative Strand

The qualitative inquiry component of the investigation used purposive criterion-based stratified sampling (Mertens, 2005) and semi-structured task-based interview methods (Seidman, 1998) that mirror the quantitative self-efficacy and calibration procedures. See Appendix D for the initial interview protocol. The criteria for interview sampling were partially derived from students' performance on the quantitative measures in an initial midterm examination. Secondary mathematics education majors were ranked based on composite measures of self-efficacy and calibration and purposely sampled in an effort to seek maximum variation (Patton, 2002) in the interview data. The 45-60 minute task-based interview protocol (Appendix D) called for participants to rate their self-efficacy to complete 5 to 7 tasks related to their course, after which participants were asked to complete three tasks ranging in difficulty level. Analysis of the task-based interview data included thematic coding (Patton, 2002) using the hypothesized model for performance and the four sources of self-efficacy as initial codes, with revised codes emerging during data analysis.

Model of the Mixed Methods Design

The research design included what Creswell (2003) refers to as the *concurrent triangulation strategy* for mixed methods research. This strategy "is selected as the model when a researcher uses two different methods to confirm, cross-validate, or corroborate findings within a single study" (Creswell, p. 217). The concurrent triangulation strategy is a traditional way to incorporate quantitiative and qualitative data sources, benefits from the potential to off-set limitations inherent in each approach, and involves integration of results from each method during the interpretation phase. Figure 9 summarizes this strategy in Creswell's diagram form. The capitalized letters in the quantitative strand, "QUAN", indicates the relative emphasis on the quantitative strand of the inquiry in relation to the qualitative strand, the "+" indicates concurrent data collection in the two strands, and the vertical arrows indicate passage of temporal order in the design.

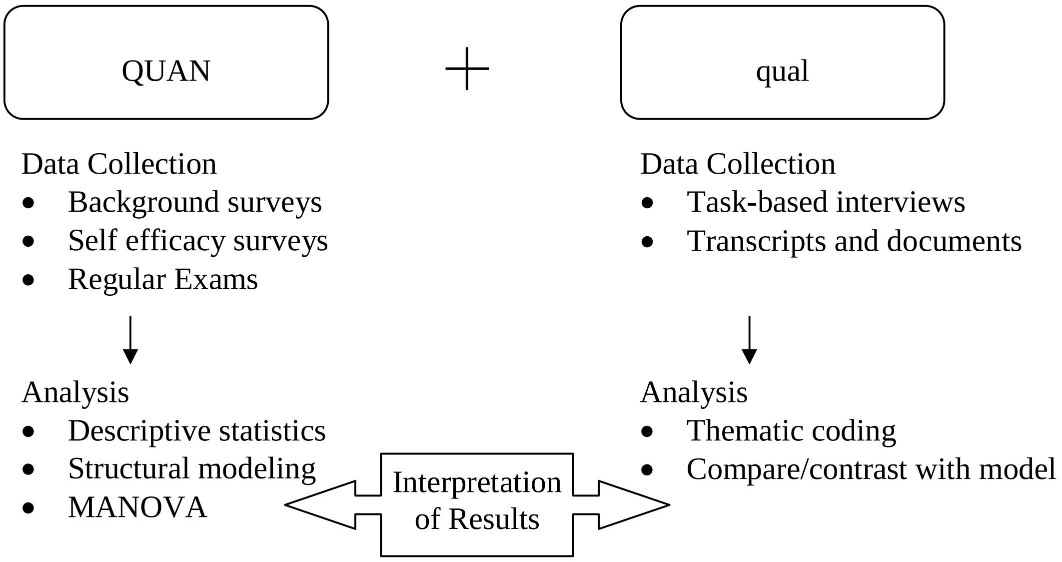

Figure 9. Model of the mixed methods design, concurrent triangulation strategy.

Assumptions of the Research Design

In addition to assumptions inherent in the theoretical perspective and social cognitive view of learning described earlier, several noteworthy suppositions are implicit in the research design:

1) Advanced mathematics students can assign numeric values to feelings of confidence toward specific tasks in their courses.
2) Latent psychological variables such as self-efficacy and calibration can be approximated by observable data (underlies the structural model).
3) Participating students’ processes for evaluating mathematics self-efficacy on final exams are similar to those they report in a task-based interview.
4) Final examinations have face validity and content validity (Creswell, 2003) as measures of performance in advanced mathematics courses.

Data Collection and Analysis

The research design included data collection procedures meant to minimize interruptions to the research setting while still collecting valid data on students' self-efficacy and mathematics performance in proximity to regularly scheduled examinations. Table 11 summarizes the data collection timeline, which began during the 8th week of classes and ended the week after final exams during the 16th week of the semester. Important phases of the data collection timeline included administering informed consent procedures, collecting background data from all participants, administering self-efficacy surveys to participants during midterm exams in seven classes, recruiting and interviewing 10 participants, and working with instructors to develop and administer final exam self-efficacy surveys in a dozen mathematics sections during the final week of the semester.

Table 11.

Timeline for Quantitative and Qualitative Data Collection

Quantitative Strand	Qualitative Strand	During
• Informed consent in 7 classes • Background surveys in 7 classes • Self-efficacy surveys during midterm exams in 7 classes		Weeks 8-9
• Preliminary analyses of quantitative surveys and performance on midterms	• Select stratified sample for task-based interviews • Recruit interview participants	Weeks 8-11
• Informed consent and background surveys in remaining classes		Weeks 11-13
	• Conduct task-based interviews	Weeks 12-13
• Create final exam survey instruments with instructors		Weeks 14-15
• Administer final exam surveys to all sections and copy student work		Week 16

Instruments

Self-efficacy instruments were developed in conjunction with the participating instructors during the week prior to the regular administration of exams. Once provided with an advance copy of an upcoming exam with instructor ratings of item difficulty, the researcher selected seven tasks from the exam to construct a self-efficacy survey that was representative of the content and difficulty of the exam. The instructor then checked the survey for face validity, and the potentially revised survey was administered to students in the few minutes just prior to the exam. See Appendix C for the self-efficacy surveys.

In addition to the construction and administration of self-efficacy surveys, the quantitative strand included data collection regarding students' subsequent performance on mathematics exams. Performance on exam items was measured primarily through the researchers' dichotomous ordinal scoring of each students' response (0 = incorrect, 5 = correct) using answer keys provided by the instructor. Estimates of the percentage of inter-rater agreement between the researchers' ratings and the instructors' ratings on a randomly selected sample of students' exams helped to indicate reliability in the performance scores.

Though the study included 12 sections of advanced mathematics courses, there were eight essentially different final exams administered by instructors. That is, four pairs of sections were taught by the same instructor and received very similar final exams. The eight unique exams included between 14 and 49 graded items (M = 25.5, SD = 11.1) each, and study participants were evaluated by a mean of 23.5 (SD = 9.3) final exam items. Instructors administered final exams during university-scheduled 2 ½ hour time periods. Some instructors allowed students to continue working for up to an additional 30 minutes, but no instructors reported a large number of students failing to finish the final exam in the time period allotted. All of the exam items had an open-response format and the instructors subsequently graded the exams for partial-credit as part of regular assessment in the classes. Photocopies of the instructors' graded final exams were collected for 7 of the 8 final exams (all except abstract algebra).

Although both the quantitative research questions and the qualitative research question aimed to provide insight into the relationships between self-efficacy, calibration, and performance in advanced mathematics exams, the analysis of data differed substantially by the type of research question and the nature of data. The six quantitative research questions split into two basic types—those relating to the hypothesized model for performance in advanced mathematics (Q1-Q4), and those pertaining to potential differences in endogenous variables by gender and test-item difficulty (Q5 & Q6). Task-based interview data served as the resource for addressing the qualitative research question (Q7). The researcher analyzed the interview data using thematic coding and descriptive vignettes (Patton, 2002).

Analysis of Quantitative Data

The first four research questions (Q1- Q4) addressed effects posited by the structural path model of performance in advanced mathematics. Structural modeling was conducted using *R*, the open source implementation of S-Plus, and relied heavily upon structural model fitting routines in the package *sem* (Fox, 2009). The *sem* implementation of structural equation modeling used similar specification conventions and produced similar statistical reports as the structural modeling program *LISREL* (Joreskog & Sorbom, 2008). Consequently, the modeling procedures followed guidelines developed by Mels (2006) and Byrne (1998) and reporting of structural modeling results followed guidelines by Schrieber (2008).

While structural equation modeling can be used for a variety of purposes, including confirmatory factor analysis and simple regression analyses, the data analysis procedures followed a seven stage process outlined by Hair and colleagues (1998, p. 592-616). The initial three of Hair and colleagues' seven stages have been described in the review of literature and theoretical perspective: (1) developing a theoretically based model, (2) constructing a path diagram corresponding to causal relationships, and (3) converting the path diagram into structural and measurement models. The remaining four stages are reported in the structural modeling results at the end of Chapter IV, including (4) choosing the input matrix type and estimating the proposed model, (5) assessing the identification of the structural model, (6) evaluating goodness-of-fit criteria, and (7) interpreting and modifying the model.

The final of the preceding stages suggests the possibility of analyzing alternative specifications of models for the data. However, “any application of structural equation modeling should have a steadfast reliance on a theoretically based foundation for the proposed model and any modifications” (Hair et al., 1998, p. 616). For this reason, the structural modeling procedures did not include consideration of alternate constructs, but rather focused on the removal of hypothesized directional effects of indicators or constructs not supported by the correlation matrix (Suhr, 2008).

The reporting of findings followed Stage, Carter, and Nora’s (2004) suggestions for path analytic research designs: (1) explicit model construction based on literature, (2) discussion of all preliminary analyses, (3) report of fit indices for all examined models, (4) illustration of final model, (5) discussion of findings in the context of previous research.

Research questions Q5 and Q6 relate to potential differences in endogenous variables in the structural diagram (e.g., self-efficacy, calibration, mathematics performance) by gender and item difficulty, respectively. Because of the inter-correlated nature of the endogenous variables, and the fact that both gender and item difficulty are considered to be categorical, multivariate analysis of variance (MANOVA) procedures were appropriate (Stevens, 1996). Reporting of tests for significant differences by gender and item difficulty was expected to follow Chen (2003), including Wilk's λ-values and *p*-values of the MANOVA tests, as well as means and standard deviations of the endogenous variables at each level of the categorical variables and post-hoc tests for differences by level of the categorical variable using Tukey's honestly significant differences criterion. Moreover, the MANOVA analysis included checks for violations of the statistical assumptions of multivariate regression, including independence of observations, multidimensional normality of the dependent variables, and approximately equal covariance matrices of groups within each variable. In particular, Box's *M* test (Stevens, 1996) was used to test for approximately equal covariance structures across levels of the categorical variables (gender and item difficulty).

Two quantitative analyses served as indications of reliability in participants' self-efficacy ratings. First, the internal consistency of students' responses was assessed using the Cronbach's alpha statistic. Second, a portion of the self-efficacy surveys included a single pair of parallel items so that, with a sufficient number of such one-point measurements of split-half reliability in the students' responses, a bivariate correlation between students self-efficacy on parallel tasks could give additional indications of reliability. Finally, the qualitative analysis of students' responses triangulated indications of reliability derived from the quantitative analyses.

In addition to the data analysis specifically designed to address the research question, it was also important to develop a richly descriptive account of the participants and the observed data. This included measures of shape, central tendency, and spread for all biographical, exogenous, and endogenous variables, including counts, means, standard deviations, tests for skewness and normality, box plots, and histograms to describe the distribution of responses in each variable. Moreover, basic bivariate associations were investigated, including correlations and cross-sectional split-plots.

Following the advice of Dr. Susan Hutchinson (2009, Personal Communication), an expert on structural equation modeling, the researcher followed some initial steps to verify the viability of the level-based indicators for the structural modeling. These included inspection of measures of internal consistency within the three constructs with a criteria of least Cronbach's $\alpha = 0.6$, significant inter-item correlations within indicators, and factor analysis of the indicators within a construct for significant factor loadings of at least $\beta = 0.4$ in a single factor principal component analysis. If the observed correlations met these criteria, the statistics served as evidence to support inclusion of the indicators and associated constructs in future structural model estimates.

Analysis of Qualitative Data

Analysis of the qualitative task-based interview data included thematic coding (Patton, 2002) of interview transcripts and artifacts from the participants' problem-solving efforts. In addition, quantitative calibration, self-efficacy, and performance results of interview participants on the midterm and final exams were integrated into the qualitative coding procedures, both as a cross-validation technique (Creswell, 2003) and as a form of data triangulation (Guion, 2002).

The qualitative data analysis process also included further review of literature as themes emerged from the data. That is, the qualitative data analysis and review of relevant literature were viewed as a cyclical process, with results from both efforts informing the other. After data analysis was completed in each of the two research strands—quantitative and qualitative—results were compared and contrasted in the interpretation of findings. It is in this data interpretation phase that the power of mixed methods research is most widely accepted (Creswell, 2003), and the goal was that the qualitative findings would help clarify, contextualize, and extend the statistical trends identified through the structural equation modeling and MANOVA techniques.

Reliability and Validity in the Quantitative Strand

In the post-positivist perspective on the quality of quantitative data collection and analysis, it is important to consider the reliability and validity of measures, procedures, and constructs in the research design. Many of the data collection procedures were designed to support claims of validity, including (1) the use of self-efficacy and calibration protocols that mirrored procedures used in related literature and two pilot studies at the research site, (2) repeated measures of self-efficacy, calibration, and performance for students in 7 of 12 class sections, (3) checks for response bias on self-efficacy surveys, (4) a background survey design based on analysis of registration data at the research site in the two previous spring semesters, and (5) analysis procedures that help to evaluate the statistical power of findings from structural equation models and MANOVA techniques. Nonetheless, all research designs include trade-offs, and the cross-sectional nature of the research design and inclusion of authentic assessment tasks introduces variation in students' responses by instructor and class section which may threaten the reliability of self-efficacy, calibration, and performance measures.

As summarized in the review of literature, self-efficacy and calibration are domain and task-specific psychological constructs and, in the case of advanced mathematics performance, are likely multi-dimensional in nature. As a result, standard measures of internal consistency, such as Cronbach's alpha and Kuder-Richardson formulae (Mertens, 2005), may provide limited information on the reliability of self-efficacy and calibration data. However, O'Connor's (1989) review of calibration research suggests test-retest reliability of self-efficacy and calibration measures is typically relatively high, with reported test-retest reliability coefficients ranging from .72 to .85. Chen (2003) reports parallel-task internal consistency coefficients to be .89 among her sample of middle school students, and composite measures of self-efficacy have been reported to be highly reliable, with Cronbach's alpha values ranging from .86 to .92 (Pajares & Graham, 1999). Thus, three procedures regarding students' self-efficacy ratings—measures of internal consistency, analysis of parallel-item reliability, and qualitative coding of self-efficacy ratings during task-based interviews—converged to provide complimentary information on the holistic reliability of self-efficacy ratings.

Trustworthiness in the Qualitative Strand

Interpretations of qualitative findings are often considered in light of descriptions of research choices that affect credibility, transferability, dependability, authenticity, and confirmability (Mertens, 2005) in the research. Credibility is akin to internal validity in quantitative research and can be supported through prolonged engagement in the research field, peer debriefing, member checks, and triangulation measures (Guion, 2002). The research design specifically addressed credibility through the cross-validating prospects of mixed methods, multiple forms of data, and parallel quantitative survey and qualitative task-based interview protocols. Moreover, credibility was supported by theoretical triangulation (Patton, 2002) in the form of converging perspectives offered by the social cognitive and cognitive information processing views of calibration.

Several choices in the research design were aimed at strengthening the transferability and dependability of qualitative findings. In particular, purposive sampling of participants to using stratified groups, together with rich, thick descriptions (Glesne, 2006) of the research setting and participants' approaches to problem-solving tasks was used to help readers evaluate the potential utility of any findings in contexts outside of the research site. Moreover, a confirmability audit (Mertens, 2005) is provided in the report, including a thorough description of all stages of data collection and analysis that led to findings in the qualitative strand of the inquiry and a full list of the codes used in the thematic analysis of interviews (Appendix F).

Authenticity, or presenting a balanced view of all perspectives, values and beliefs of participants in a research setting, is a critical element of qualitative reporting. This was partly addressed in the research design through the open-ended, think-aloud, task-based interview protocol, but was also addressed through qualitative inquiry strategies that emphasized the variety of students' experiences of mathematics self-efficacy, calibration, and exam performance while avoiding tendencies to report general quantitative and qualitative trends as universalities or hard-and-fast rules. The aim of the qualitative strand of the inquiry was to include a sense of the wide personal variation that was more difficult to get from the statistical findings, and fairness to the multiple perspectives of participants was an important value informing the design, analysis, and reporting.

Limitations and Delimitations in the Research Design

While adding to limited educational research into the mathematics self-efficacy and calibration of college students in general, and preservice mathematics teachers' in particular, the research design reflected many methodological choices regarding scope, procedures, and criteria for drawing conclusions. These choices have narrowing consequences, or *delimitations* (Creswell, 2003), on the research value of findings flowing from the research design. One such delimitation was the choice of restricting data collection and analysis of participating students' mathematics performance to traditional in-class examinations. Alternative measures of advanced mathematics achievement such as writing assignments, projects, take-home tests, and laboratory reports have been advocated by a number of mathematicians (Rosenthal, 1995) and were used by some instructors at the research site, but the measures of mathematics self-efficacy and performance focused only on traditional in-class examinations.

Additional delimitations in the research design included restricting qualitative inquiry to a relatively small number (10) of interviews in seven mathematics classes at a single university. The trustworthiness of the narrative in the qualitative strand could have benefited from additional data sources such as classroom observations or interviews with faculty or member-checking of results with the interview participants. While the limited qualitative data flowed from the comparatively large quantitative component in the research, they also limited the potential generalizability of findings from the qualitative inquiry. However, generalization of characteristics from a sample to a target population is not necessarily an aim of qualitative inquiry (Patton, 2002), and the ultimate value of the qualitative strand derived from the transferability and trustworthiness (Mertens, 2005) of the holistic descriptions and thematic analysis which supported, contrasted, and added context to the broad statistical trends identified in the quantitative strand.

The quantitative strand of the design focused on a limited number of potentially significant intervening variables (i.e., gender, item difficulty, high school mathematics achievement, and item difficulty). Some intervening variables identified in mathematics self-efficacy research were not included in the research design, including general intelligence, math anxiety, academic level, goal structures, self-concept, learning disability, and socio-economic status (see Pajares & Urdan, 2006). Moreover, potential classroom-level effects of instruction on students' mathematics self-efficacy were not included in the research design, largely because of the limited sample size and small number of participating instructors in the design. Finally, while the literature review suggested that mathematics self-efficacy develops over time with experience in the domain, the cross-sectional nature of the research design did not allow for inferences into the longitudinal development of mathematics self-efficacy.

Several elements of the research design introduced limitations (Creswell, 2003), or potential weaknesses, in the study. First, the cross-sectional survey design did not include experimental control, time order, or manipulation, all of which are required to make claims about causality (Frankfort-Nachmias & Nachmias, 2000). That is, any statistical effects identified in the quantitative component of the design can only be used to explain relationships among variables and cannot provide evidence of causation. The lack of measures to control intervening variables such as instruction, while supporting the naturalistic case-study inquiry, also introduced threats to the internal validity of the research design, described by Colosi (1997) as the amount of evidence to support directional effects. For instance, there was potential for history effects (Frankfort-Nachmias & Nachmias, 2000) in the form of variation in instructional practices across classes at the research site, and instrumentation effects in the form of differences in item difficulty and self-efficacy prompts presented to participants in different courses. The research design included a snapshot (Creswell, 2007) of students' academic behaviors and beliefs at a single point in time, and the limited study duration permits the possibility that observed relationships may have changed over time.

There were also threats to the external validity of the research design in terms of generalizing findings and themes to populations outside of the participants at the local research site. While all students taking mathematics content courses at the research site which offer traditional in-class examinations had the opportunity to participate in the study, the study was limited to a single university in one academic semester and thus only included students and instructors involved in the mathematics courses for prospective secondary mathematics teachers in a narrow time-frame and specific context.

Enrolled students were not randomly assigned to mathematics courses. The instructors and research site were not randomly sampled from the larger population of faculty members and universities that prepare secondary mathematics teachers in the United States. Readers of this report are encouraged to consider any findings and implications in light of these limitations and draw on the description of the research setting, participants, and methodology to evaluate the transferability of any findings to other settings containing secondary mathematics teachers and other students enrolled in advanced mathematics courses.

Research Timeline

Table 12 outlines the research timeline for the study, including the completed pilot studies and phases for data collection, analysis, and narrative summary of findings.

Table 12.

Summary of the Research Timeline, Fall 2007- Spring 2010

Research Phase	Progress	Completion
Pilot Study (College Algebra)	Development of survey instruments, data analysis methods, initial findings. Exploratory qualitative investigation.	Fall 2007
Pilot Study (Calculus I)	Refinement of survey instruments, data analysis. Informed consent and survey methods.	Spring 2007- Summer 2007
Dissertation Proposal	Literature review, conceptual framework, research design, methods, post-hoc analysis of pilot study data.	Fall 2008
Data Collection	Inform consent, administration of surveys, observation of classrooms, task-based interviews.	Spring 2009
Data Analysis	Statistical modeling, transcription and coding of interviews, analysis of observation data	Summer 2009
Dissertation Completion	Results, findings and discussion	Fall 2009
Defense	Completion of dissertation, dissemination of findings, dissertation defense	Early Spring 2010

CHAPTER IV

RESULTS

The purpose of this chapter is to summarize the results of the cross-sectional study described in the previous chapter. Building on the descriptive accounts of the participants and sample in the methodology chapter, the results include descriptive and inferential statistics arising from the process of addressing the seven research questions and five hypotheses. Results pertaining to the six quantitative research questions, given below as Q1 – Q6, arose from structural equation modeling and analysis of variance techniques. These quantitative findings were contextualized by findings regarding the single qualitative research question (Q7).

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects on participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

Q7 In what ways do prospective secondary mathematics teachers' mathematical problem-solving compare and contrast with the hypothesized relationships between self-efficacy, calibration, and performance in advanced mathematics?

Inherent in the first four research questions is the *a priori* structural model relating a single exogenous latent construct (high school mathematics achievement) to four endogenous latent variables (Figure 10). The theoretical rationale for the directional effects, which is central to the validity of the modeling procedures, is described in the review of literature in Chapter II.

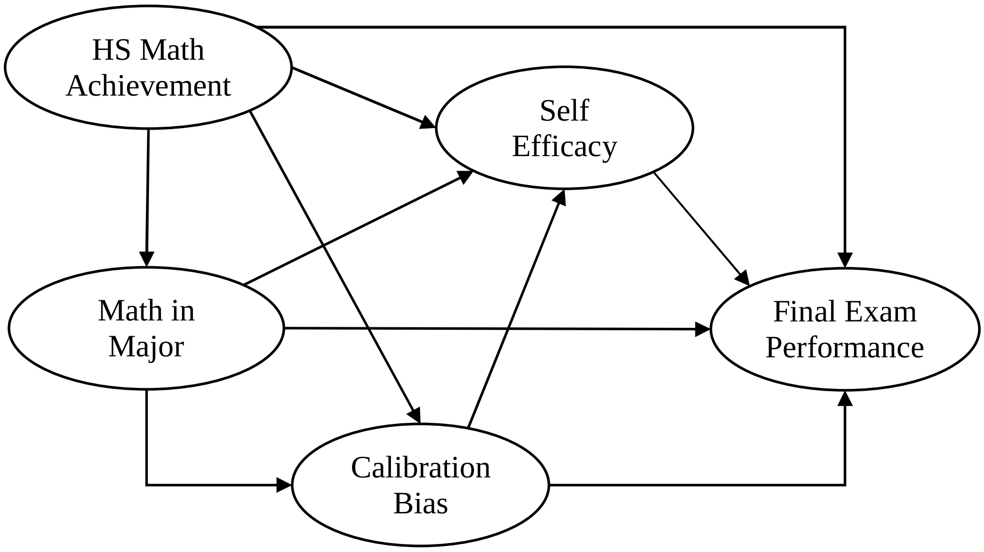

Figure 10. Hypothesized structural path model for advanced mathematics performance.

Forthcoming sections in this chapter detail results of statistical analyses aimed at addressing the quantitative research questions and hypotheses. Initially, the findings focus on descriptive summaries of contextual and background information on the participants. Then, the narrative presents results of analyses of the statistical evidence regarding potential differences in the variables identified in research questions Q5 and Q6 associated with participants' gender and the difficulty of exam items. Next, the summary includes results of the structural modeling of relationships among the amount of mathematics in participants' college major, high school mathematics achievement, self-efficacy, calibration, and mathematical performance. Finally, the narrative addresses evidence supporting five qualitative themes surrounding the mathematics self-efficacy, performance and calibration of secondary mathematics majors.

Quantitative Data

In cross-sectional research, inferential statistics and modeling results are better understood in the context of the distributional characteristics of study data (Schrieber, 2008). Consequently, initial steps in the analysis of the quantitative data included descriptive summaries of indicator variables used in the structural modeling and composite measures of self-efficacy, calibration, and final exam performance.

Continuous Indicators of Latent Constructs

Table 13 includes descriptive statistics of the indicator variables for the latent constructs High School Math, Math in Major, Self-Efficacy, and Calibration Bias. Some highlights of the indicator distributions include (1) ascending means for indicators of self-efficacy and calibration by "level" with some negatively skewed self-efficacy indicators, (2) moderate-to-high ACT Math scores and self-assessments of high school mathematics performance, (3) skewed-left high school GPA scores with an apparent ceiling effect at 4.0, and (4) bimodal distribution of required mathematics credits associated with students' college majors. The following brief sections summarize each of these distributional characteristics.

Table 13.

Descriptive Summary of Indicators for High School Math Achievement, Math in Major, Self-Efficacy, and Calibration Bias

Construct	Indicator[a]	*n*	*M*	*SD*	Scale
HS Math Achievement	ACT Math	132	24.9	3.9	14 to 36
	HS GPA	133	3.4	0.6	0 to 4
	HS Self	195	4.6	1.1	0 to 7
Math in Major	Required Math	177[b]	23.0	15.2	3 to 45
Self-Efficacy	SE Level 1	195	2.9	1.3	0 to 5
	SE Level 2		3.2	1.2	
	SE Level 3		3.5	1.1	
	SE Level 4		3.7	1.1	
	SE Level 5		4.0	1.0	
	SE Level 6		4.2	1.0	
	SE Level 7		4.5	0.9	
Calibration Bias	Bias Level 1	195	-0.4	1.8	-5 to 5
	Bias Level 2		0.2	2.2	
	Bias Level 3		0.4	2.2	
	Bias Level 4		0.9	2.3	
	Bias Level 5		1.2	2.4	
	Bias Level 6		1.5	2.4	
	Bias Level 7		1.9	2.4	

Note. [a]"Level" indicator were formed by ascending within-class means. [b]Missing values for Required Math correspond to ambiguous majors (e.g., "undeclared", "pre-program"). SE = self-efficacy rating, HS GPA = high school grade point average (capped at 4.0); HS Self = self-assessment of high school mathematics performance, Required Math = number of semester mathematics credits required by declared college major.

Indicators of Self-Efficacy and Calibration Bias

As described in the methodology and evidenced in Table 13, means of the indicators for self-efficacy and calibration bias are ascending by "level." For example, a student's SE Level 1 rating indicates belief in being able to complete the mathematical tasks in which his or her classmates expressed the lowest collective rating. The indicators of self-efficacy suggested students' mean self-efficacy ranged from $M = 2.9$ to $M = 4.5$

on a scale of 0 to 5 on the sample of exam items presented during the pre-final exam surveys, with similar observed variation (standard deviations ranged from 0.9 to 1.3) at each level. Calibration indicators ranged from Level 1 (M = -0.4, SD = 1.8), which indicated significant overall underconfidence (t (194) = 3.1, $p < .01$) to Level 7, which indicated significant overall overconfidence (M = 1.9, SD = 2.4, t (194) = 11.1, $p < .001$). However, observed calibration means were significantly positive in 5 of 7 indicators at the $\alpha = .01$ criterion, suggesting general tendencies toward overconfidence in the calibration indicators.

There was some evidence that self-efficacy ratings were negatively skewed for higher-level indicators. For example, on the highest level indicator of self-efficacy, Level 7, 89% (173/195) of students marked their confidence in being able to complete the indicated task successfully as either 4 or 5 (out of 5). Of the seven indicators of self-efficacy, Levels 3 - 7 were all significantly negatively skewed at the $\alpha = .01$ criterion (skew = -0.5, -0.9, -0.9, -1.2, -1.9, kurtosis = -0.4, 0.5, 0.3, 1.6, 5.1, respectively). However, the primary purpose of the level indicators of self-efficacy was for structural modeling, which typically produces estimates with robust standard errors when the absolute values of skewness are below 2.0 and kurtosis parameters are below 7.0 (Schreiber, 2008). Having met these criteria, all indicators of self-efficacy and calibration bias were retained for the structural modeling.

The norm-referenced definition of the indicators permits the theoretical possibility that within-class means may differ from the composite means. However, observed deviations of within-class means from the overall means of indicators were less than 0.2

and were inconsistently related to class section, thus providing empirical support to the use of level indicators of self-efficacy and calibration based on within-class means.

Indicators of High School Mathematics Achievement

The data suggested study participants achieved relatively high ACT Math scores. The average ACT Math score of M = 24.9 corresponds to approximately the 79th percentile of U.S. college-bound students (ACT, 2007). Based on a large sample of data on new students at the research site (Fitchett, King, & Champion, in press), the students enrolled in the 12 participating sections entered college with an average ACT Math score about one standard deviation above their peers at the university, a difference which is statistically significant (M = 21.6, SD = 3.7, t(1,236) = 11.3, d = .9, p < .001). Graphical checks for normality (i.e., Q-Q-plots) suggested retaining the assumption that ACT Math scores were normally distributed.

Participants' self-assessments of their high school mathematics performance were in line with the relatively high mathematics achievement indicated by the distribution of ACT Math scores. Figure 11 shows the distribution of the participants' ratings for the question, "Which of the following best describes how well you did in your high school math courses?" Most students chose one of the descriptors "OK" (27%), "Good" (39%), or "Very Good" (21%), indicating moderate-to-strong self-assessments of high school mathematics performance. Graphical checks for normality suggested retaining the assumption that students' self-assessments were normally distributed.

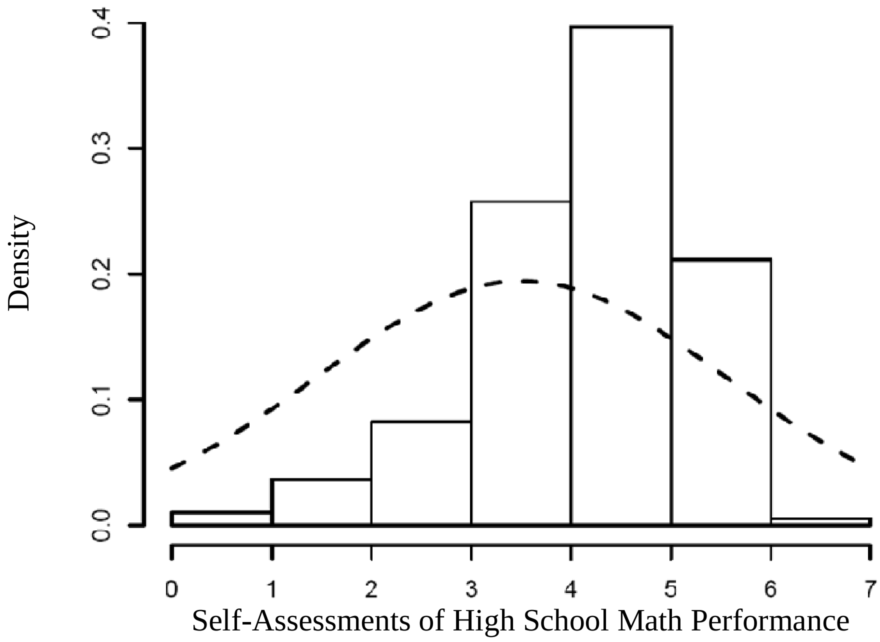

Figure 11. Distribution of students' self-assessments of their high school mathematics performance (N = 195). 1 = really bad, 2 = bad, 3 = not-so-good, 4 = ok, 5 = good, 6 = very good, 7 = excellent. For reference, the dashed curve indicates a normal curve with the sample mean and standard deviation.

Study participants' high school grade point averages were significantly skewed-left (n = 133, M = 3.4, *mdn* = 3.6, skew = -0.7, kurtosis = -0.6, $p < .001$). This is potentially attributed to an admissions policy at the research site that caps high school grade point averages at 4.0 (many secondary schools in the state award "honors" points for some classes that may lead to grade point averages above 4.0). This apparent ceiling effect is evidenced by the fact that 17% of the reported high school grade point averages were exactly 4.0. In contrast, Fitchett et al. (in press) found an approximately normal distribution of high school GPAs (M = 3.2, SD =0.4, n = 1029, skew = -0.1) among new students at the research site, with only 7% equal to 4.0. This suggests the high proportion of 4.0 GPAs in the sample may be atypical of the undergraduate population at the research site. Figure 12 shows a histogram of the high school grade point averages of study participants, including a noticeable spike at 4.0.

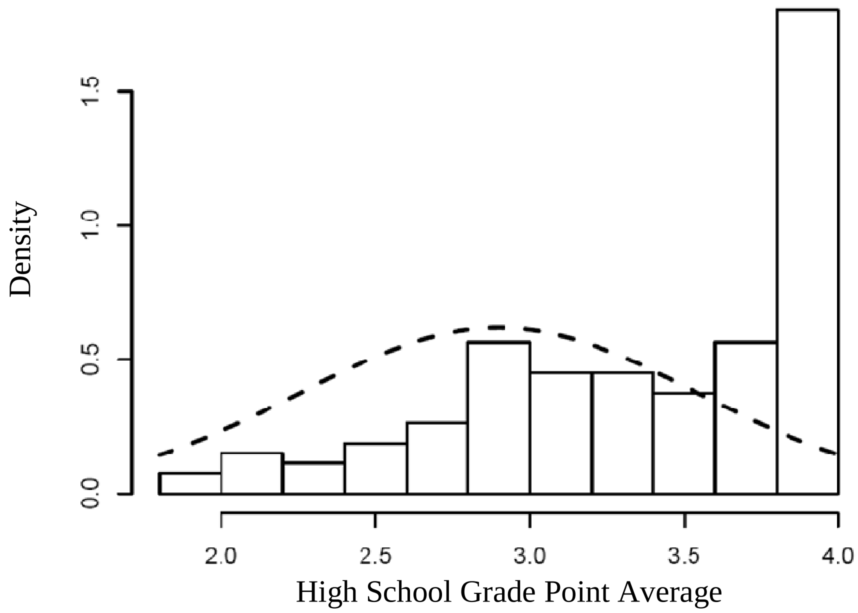

Figure 12. Distribution of students' high school grade point averages (*n* = 133) . Grade point averages are capped at 4.0 by university admissions procedures. For reference, the dashed curve indicates a normal distribution with the sample mean and standard deviation.

Required Math as an Indicator of Math in Major

The distribution of required semester mathematics credits in participants' primary college majors (Required Math) is derived from the distributions of students' majors in Table 5. The most common majors declared by study participants were Elementary Education – Mathematics (26%), Mathematics – Secondary Teaching (9%), a non-teaching Mathematics concentration (14%), and Chemistry (10%). The result of these proportions is the bimodal distribution of Required Math credits shown in Figure 13. The large proportion of Mathematics majors in the sample produced a large singular departure from normality near 40 required credits; the remaining non-mathematics-only majors formed a separate approximately normal distribution centered near 10 required mathematics credits.

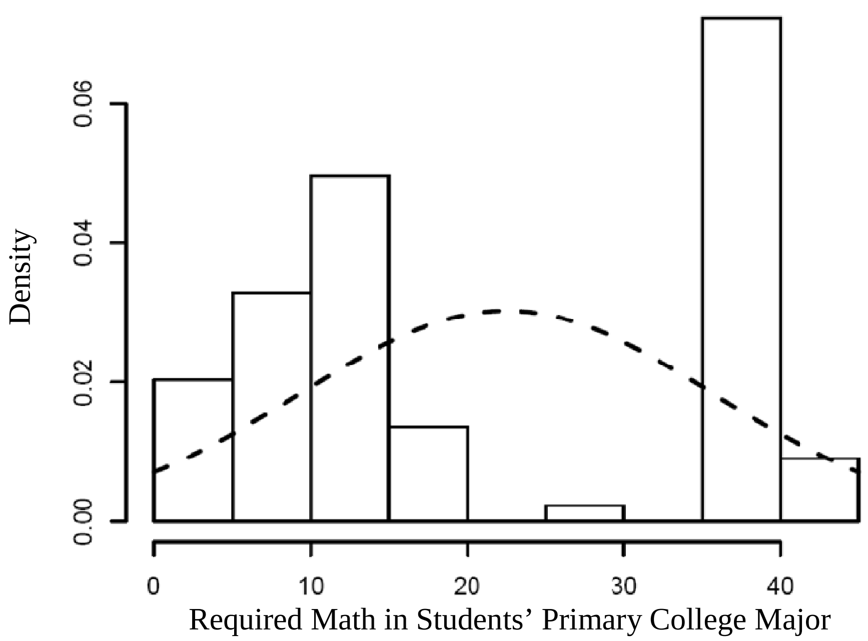

Figure 13. Distribution of the number of semester credits with prefix MATH required by students' primary declared college major (n = 177). Dashed curve indicates a normal distribution with the sample mean and standard deviation.

Indicators of Final Exam Performance

Collectively, the seven dichotomous indicators of final exam performance represent students' performance on exam items over a range of difficulties. Table 14 gives the distributions of students' work which was scored as "correct" or "incorrect" for the sampled final exam items corresponding to the seven "level" indicators. Since final exam items were sampled for level indicators based on within-class item difficulty, the level indicators are ordered so that higher levels are associated with a higher percentage of correct student responses. Performance Level 1, for example, represents students' performance on a difficult final exam item – only about one in four students (25%) correctly solved the task corresponding to this first indicator. In contrast, 83% of participants correctly solved the final exam item corresponding to the Performance Level 7 indicator. Three indicators come from final exam tasks which were correctly solved by fewer than half of students; the remaining four indicators include greater than 50% correct responses.

Table 14.

Distributions of Indicators for Final Exam Performance

Construct	Indicator	% Incorrect	% Correct
Final Exam Performance	Perf. Level 1	75	25
	Perf. Level 2	63	37
	Perf. Level 3	56	44
	Perf. Level 4	42	58
	Perf. Level 5	33	67
	Perf. Level 6	27	73
	Perf. Level 7	17	83

Note. Table entries indicate the proportion of students (N = 195) who correctly solved the corresponding final exam items. Items were randomly sampled from seven-level quantile groups based on item difficulty.

As in the analysis of self-efficacy and calibration indicators, the analysis of level indicators included checks for variation in performance across the sampled exam items that composed each indicator. These observed within-class distributions of performance indicators differed from composite distributions by up to 4%, but differences were inconsistent by section and supported retaining the assumption that indicators held similar distributions across sections.

Composite Scales of Self-Efficacy, Calibration, and Final Exam Performance

While self-efficacy, calibration bias, and final exam performance were each considered to be latent constructs measured by seven indicators, the research question which considered potential differences in these constructs (along with Math in Major) by participants' gender (Q5) called for a composite scale for each construct. Table 15 gives a descriptive summary of the scales constructed as means of the seven level indicators within each construct and includes the observed Cronbach's α values corresponding to the scales.

Table 15.

Descriptive Summary of Self-Efficacy, Calibration Bias, and Final Exam Performance

Composite Scale	*M*	*SD*	Range	Cronbach's α[a]
Self-Efficacy	3.7	0.7	0 to 5	.77
Calibration Bias	0.8	1.1	-5 to 5	.53
Final Exam Performance	2.9	1.3	0 to 5	.65[b]

Note. Composite scales represent means of the seven level indicators of each construct. [a]Alpha values indicate the expected correlation of two scales constructed by sampling seven items from within each construct (Bland & Altman, 1997). [b]Final exam indicators were dichotomous, so α was calculated as the Kruder-Richardson KR-20 coefficient.

The mean composite self-efficacy rating (M = 3.7) was significantly larger than the mean performance score (M =2.9, t (388) = 8.5, d = 1.0, p < .001), with the 95% confidence interval for the difference being 0.7 to 1.1. This suggested a significant overall trend toward calibration bias in the form of overconfidence. Interestingly, the observed calibration bias mean (M =0.8) fell within the 95% confidence interval of the difference between self-efficacy and performance, despite the fact that the final exam performance indicators were not matched to the final exam items used in the self-efficacy and calibration measures.

Cronbach's α values given in Table 15 were interpreted directly so that each α value gives the expected correlation between the observed scale and a second theoretical scale constructed by random sampling seven items from the sample space of items which could be used to measure the latent construct (Bland & Altman, 1997). For example, had seven different final exam items been chosen for the self-efficacy surveys, and a separate calibration bias score calculated, the correlation between this new calibration measure and the observed calibration measure would be expected to be moderate (α =.53). However, the correlation between the second composite self-efficacy scale and the observed self-efficacy scale would be expected to be high (α =.77).

Reliability and Validity of Self-Efficacy and Final Exam Performance Indicators

One of the strengths of structural modeling is the estimation of measurement error, so that assessing the extent to which indicators of a construct reflect consistent measurement of a single construct was built into the analysis of the measurement model and is consequently reported in the structural modeling results. In addition to this factor-analytic approach, several efforts were taken to estimate the reliability of self-efficacy ratings and final exam performance scores. These include internal consistency, parallel-task reliability for a sample of self-efficacy survey ratings, and inter-rater reliability of final exam performance scoring.

As a measure of internal consistency among ratings, Cronbach's α is often used indirectly to assess the reliability of an instrument which contains multiple items designed to measure a single construct (Hair et al., 1998). From this perspective, the composite measures of calibration bias (α = .53) and final exam performance (α = .65) fail to meet the traditional benchmark for adequate reliability of a unidimensional construct (Cronbach's α >.7, Bland & Altman, 1997). Since calibration bias is calculated as the difference of self-efficacy ratings and performance scores, one possible source for the relatively low reliability of the calibration bias scale is the cumulative variation due to measurement errors from both final exam performance and self-efficacy. Moreover, the value of Cronbach's α as an indirect measure of reliability is reduced to some extent by the relatively small sample (n =7) of items contributing to the scales because Cronbach's α is an increasing function of the number of items comprising a scale (Revelle, 2009).

A more direct measure of reliability in the self-efficacy ratings was the students' ratings on parallel tasks included on each of the self-efficacy surveys administered to n = 131 students in the minutes just before the mid-term exams in seven of the participating sections. Each of these midterm self-efficacy surveys contained a single pair of parallel tasks, and reliability would be strengthened if students' self-efficacy ratings on the separate tasks were similar. This was confirmed by nearly identical mean ratings on the first and second of the parallel tasks presented to students (M = 4.08, SD = .90, and M = 4.06, SD = .91, respectively). In 73% (96/131) of the cases, students' ratings on the two tasks were identical, and in 95% (125/131) of the cases the two ratings were within 1.The split-half correlation between the two ratings was high (r = .71).

Students likely completed the self-efficacy surveys under the belief that the correctness of their final exam solutions would be determined by their instructors. Thus, the validity of the final exam indicators would be weakened by potential differences between the instructors' scoring of final exams and the researchers' dichotomous scoring of the exams. If the observed differences between the researcher scores and the instructors’ grades were proportionally small, it would strengthen the validity of the final exam scoring and would also indicate reliability in the scoring of final exam data.

Instructors’ graded final exams were available for 7 of the 8 final exams (all except abstract algebra), and from this pool, a random sample of n = 70 students’ final exams were selected for inter-rater comparison. The sample size of 70 was chosen so that the statistical power to detect significant agreement between ratings with 95% confidence was approximately 90% (Sim & Wright, 2005). Collectively, the sample yielded 1,655 ratings which are summarized in Table 16. To achieve comparable scales, instructors’ partial-credit scorings of items were converted to a dichotomous scale using an “all-or-nothing” rule—if the instructor scored a students’ performance on an item as anything less than full credit, the item was entered as incorrect.

Table 16.

Percentages of Final Exam Item Scores by Instructors and the Researcher

		Instructor Rating		
		% Incorrect	% Correct	Total
Researcher Rating	% Incorrect	40	4	45
	% Correct	7	49	56
	Total	47	53	100

Note. Entries are percentages of the 1,655 total ratings from a random sample of 70 final exams.

The inter-rater agreement comparison in Table 16 shows agreement in 89% (1473/1655) of the sampled ratings. The distribution of ratings correspond to an inter-rater reliability coefficient of κ = .83 (κ_{max} = .92), which suggests “almost perfect” agreement (Sim & Wright, 2005, p. 264). Though disagreement between instructor and researcher ratings was relatively rare, the observed percentage of ratings in which the researcher scored students' work "correct" while the instructors' grade was classified as "incorrect" was slightly larger than the reverse (7% compared to 4%). This may have been due to the strict "all-or-nothing" criterion for converting instructors' partial credit grading schemes to dichotomous scores. If instructors independently graded items on a dichotomous scale, their ratings may have differed slightly from the post-hoc "all-or-nothing" classifications.

Differences by Gender

The fifth research question (Q5), given below, asked about potential differences in the composite indicators of self-efficacy, calibration, math in major, and final exam performance associated with gender. The hypothesis (H5) included an expectation that there would be no differences in self-efficacy, calibration, or final exam performance by students' gender, but that males would have more required mathematics credits in their majors than females.

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

H5 There will be no significant difference in self-efficacy, performance, or calibration by gender. There will be significant differences in the amount of mathematics in participants' college major by gender, with males on average choosing college majors with more required mathematics courses.

Since the explanatory variable (gender) is dichotomous and the response variables are intercorrelated continuous scales, the appropriate test for differences by gender was a one-factor, between subjects, multivariate analysis of variance (MANOVA) (Grice & Iwasaki, 2007). The MANOVA analysis included (1) checks on the assumptions of MANOVA, (2) an omnibus test for differences between the male and female groups on linear combinations of the responses variables, and (3) evaluation of alternative and trimmed models.

Checks on Assumptions of MANOVA

Interpretations of MANOVA results can be limited by departures of the data from the statistical assumptions underlying the techniques. Garson (2006) identifies several assumptions which can affect the statistical power of MANOVA analyses, including (1) independent observations, (2) approximately equal group sizes, (3) adequate sample size, (4) randomly distributed residuals, (5) homogeneity of variance and covariance matrices (homoscedasticity), and (6) multivariate normality. The first three assumptions were met by the research design, so the checks on the assumptions underlying the MANOVA tests focused on evaluating model residuals, homoscedasticity, and multivariate normality.

Evaluation of whether the data met the assumptions of MANOVA was particularly important in the case of the omnibus test for differences by gender. Graphical inspection of model residuals suggested no substantial departures from normally distributed errors. Bartlett's K^2 and Brown-Forsyth's F tests for homogeneity of variances across the gender groups supported retaining the null hypothesis of approximately equal variances in the response variables. Box's M test for approximately equal covariances in the response variables (Stevens, 1996) failed to provide evidence of unequal covariances (F (10, 144321) = 1.3, p = .20). Finally, a graphical check of multivariate normality using a Q-Q-plot of the generalized distance (De Maesschalck, Jouan-Rimbaud, & Massart, 2000) of the data points from the observed centroid supported retaining the assumption of multivariate normality. In summary, the analysis supported the assumptions underlying the omnibus MANOVA test for differences by gender.

Omnibus Test for Differences

The composite one-factor model for differences in Required Math or the composite scales of math in major, self-efficacy, calibration, and final exam performance by gender was not significant (Wilk's $\Lambda = .97$, F (4, 172) = 1.3, $p = .27$), suggesting insufficient evidence to support differences in the response variables by gender. A post-hoc analysis suggested the mean Required Math for females ($M = 25.2$, $SD = 14.7$) was significantly higher than that of males ($M = 20.6$, $SD = 15.4$, t (54) = 5.4, $d = .29$, $p < .01$). The statistical power of this observed difference is potentially weakened by departure of the Required Math distribution from normality (i.e., both marginal distributions were bimodal). Nonetheless, histograms of the male and female distributions, given in Figure 14 supported small differences in the number of required mathematics credits in favor of female students. This observed difference was likely due to a small difference in the percentage of mathematics majors by gender (χ^2 (1, $N = 195$) = 3.9, $p < .05$). That is, while 44% (43/97) of female participants were mathematics majors, just 31% (30/98) of male participants were mathematics majors.

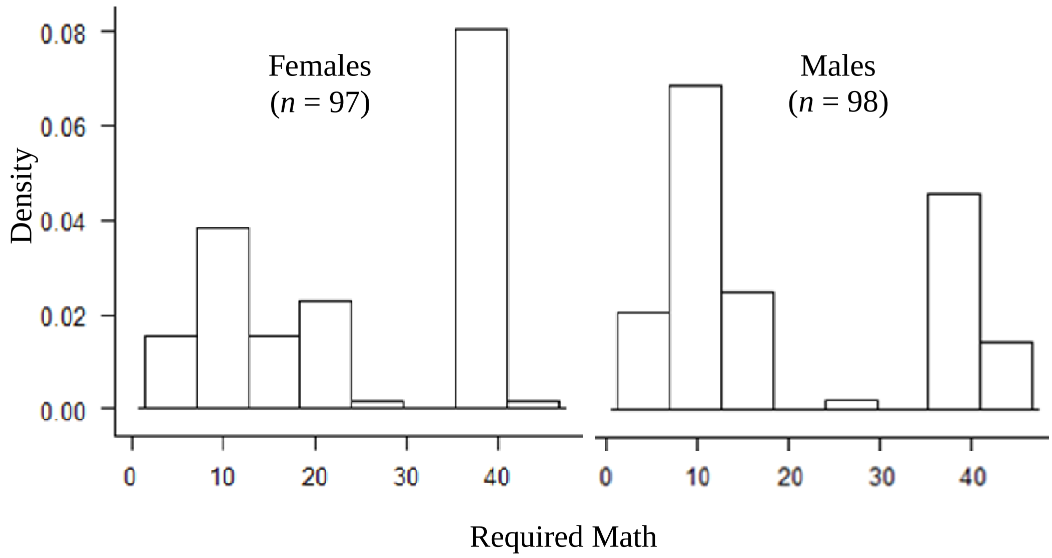

Figure 14. Marginal distributions of Required Math by gender. The histograms suggest females chose majors with more Required Math than males.

Evaluation of Alternative and Trimmed Models

The analysis of potential differences by gender included consideration of several alternative models for effects of gender on math in major, self-efficacy, calibration, and final exam performance. The alternative specifications included replacing the measure of calibration bias by calibration accuracy, replacing final exam performance by composite scales from alternate samples of exam items, and testing all 16 possible trimmed subsets of response variables (e.g., dropping final exam performance). With the exception of the highly restricted model positing direct effects of gender on math in major (equivalent to the reported post-hoc *t*-test), none of the alternative models reached significance. Especially in light of the relatively large number of degrees of freedom which made MANOVA sensitive to small differences due to gender, the results suggested very limited support for any potential differences in self-efficacy, calibration, or final exam performance attributable to participants' gender.

Differences by Item Difficulty

The sixth research question, given below, looked for differences in self-efficacy ratings and calibration scores associated with *item difficulty* – the mean student-performance on final exam items matched to self-efficacy survey items. Based on the literature review, the expectation (H6) was that there would be no difference in self-efficacy ratings associated with the difficulty of the items, but there would be a tendency toward overconfidence on survey items with increased item difficulty.

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

H6 There will be no significant difference in self-efficacy by item difficulty. There will be a significant difference in calibration by item difficulty, with a tendency toward overconfidence on more difficult exam items.

Similar to the indicators of final exam performance, the difficulty of the final exam items given in the self-efficacy surveys was determined by the reverse rank-ordering of students' final exam performance on the items. For example, a self-efficacy survey item with an item difficulty rank of "1" corresponded to the "easiest" item on the survey because the highest percentage of students in the section correctly completed the matched final exam item. Likewise, a self-efficacy survey item with difficulty "7" would be considered the "hardest" mathematical task on the self-efficacy survey because the lowest percentage of students in the class correctly completed the corresponding final exam item.

The seven measurements of students' self-efficacy and calibration bias, respectively, across varying item difficulties constitute a type of within-subjects (repeated measures) design. That is, the goal of the analyses was to test for effects of a single categorical variable, item difficulty, on two continuous dependent variables – self-efficacy and calibration bias – while adjusting for within-subject means. The omnibus test, a one-way repeated measures MANOVA, identified significant differences among self-efficacy ratings and calibration bias scores associated with item difficulty levels (Wilk's $\Lambda = 0.11$, F (13, 182) = 109.0, $p < .001$). As in the MANOVA tests for differences associated with gender, checks on the marginal distributions and covariance structures suggested the data met the assumptions of MANOVA. Importantly, though, there was limited evidence to support intercorrelation between self-efficacy and calibration bias ($r = .10$, $p = .16$), so subsequent analyses were conducted using separate one-way repeated measures analysis of variance (ANOVA) methods.

The data supported the hypothesis of significant differences in calibration bias associated with item difficulty, F (6, 1164) = 14.9, $p < .001$, but the observed differences did not support the hypothesis that students' calibration bias tended toward overconfidence with increasing item difficulty. Instead, as evidenced in Table 17, means of calibration bias scores generally decreased from overconfidence on the least difficult items (e.g., mean bias = +1.6 for Level 1 difficulty) toward near perfect calibration on the most difficult items (e.g., mean bias = 0.0 for Level 7 difficulty). Moreover, in contrast to the hypothesis, there were significant differences in self-efficacy ratings associated with item difficulty, F (6,1164) = 36.6, $p < .001$, with increasing item difficulty typically associated with decreasing self-efficacy ratings.

Table 17.

Self-Efficacy and Calibration Bias by Item Difficulty Level

	Self-Efficacy Rating		Calibration Bias Scores	
Item Difficulty	*M*	*SD*	*M*	*SD*
Level 1	4.3_{ab}	1.0	1.6_{a}	2.4
Level 2	4.1_{b}	1.1	1.3_{ab}	2.4
Level 3	3.7_{c}	1.2	1.0_{b}	2.1
Level 4	3.6_{cd}	1.1	0.5_{bc}	2.3
Level 5	3.8_{bcd}	1.2	1.0_{bcd}	2.3
Level 6	3.3_{e}	1.2	0.2_{ce}	2.2
Level 7	3.0_{f}	1.3	0.0_{ce}	2.0

Note. N = 195. Within categories, means with the same subscript do not differ significantly by the Tukey honestly significant difference test at α = .05. Increasing difficulty "Level" indicates lower success rates on corresponding final exam items.

Post-hoc comparisons of means by item difficulty level using Tukey's honestly significant difference (HSD) criterion suggested that differences in self-efficacy means were typically found for items separated by two or more difficulty levels. For example, the observed self-efficacy on items of moderate to high difficulty (Levels 3-7) differed significantly from self-efficacy on the least difficult exam item (Level 1), although the difference between means of Level 1 and Level 2 self-efficacy ratings was not significant (p = .05). In fact, the analysis identified significant differences in self-efficacy means for items separated by at least two levels of difficulty in 87% (13/15) of the possible cases.

Table 17 also reports Tukey HSD comparisons of calibration bias means, which indicated overall tendencies toward decreasing calibration bias associated with increasing item difficulty. Though mean calibration bias on the least-difficult exam items (Levels 1

& 2) were significantly greater than calibration bias on the most-difficult exam items (Levels 6 and 7), differences in means across moderate difficulty levels were inconsistent. That is, there was a general trend was toward reduced overconfidence on more difficult exam items, but calibration bias means were statistically similar for moderately difficult exam items (Levels 3-5).

In the lexicon of structural modeling, the results to this point have addressed *manifest* (observed) variables in the study. Following a descriptive account of the data, the analysis found no significant differences in composite self-efficacy, calibration bias, or final exam performance associated with participants gender (Q5), but did identify slightly greater required mathematics requirements associated with the declared majors of female participants. Then, differences were identified in self-efficacy and calibration bias by the difficulty of exam items (Q6), including trends toward decreased self-efficacy and reduced overconfidence on more difficult exam items. In the next sections, the focus shifts from manifest variables to structural relationships among latent constructs, including effects among high school mathematics achievement, self-efficacy, calibration bias, and final exam performance.

Structural Equation Modeling

Multiple Imputation of Missing Data

Of the 25 indicator variables used in the structural model, complete observed data were collected for 23 indicators. However, approximately 30% of the data for students' ACT Math and High School GPAs were missing from the registration data. In addition, though a declared college major was available for all participants, the corresponding Required Math indicator was labeled "NA" for 9% of participants because of “undeclared” or other majors whose required mathematics credits were ambiguous. Though the reasons for the missing data were unknown, one possible source of missing ACT Math and High School GPA data was non-traditional admission to the university, such as transferring to the university from another university or community college. Several strategies for handling the missing data were considered and are described below.

A first step in choosing a missing data strategy is the classification of missing data as one of *missing completely at random* (MCAR), *missing at random* (MAR), or *missing not at random* (MNAR) (Collins, Schafer, & Kam, 2001). Under the MCAR assumption, students with missing data could be deleted casewise from the data without introducing bias. However, casewise deletion would reduce statistical power and would involve the assumption that students with missing data do not differ from those with full data. The less restrictive assumption MAR would simply require that the missing ACT Math and High School GPA data were not missing *because* of other variables in the structural model. For example, if students with low final exam performance were embarrassed and thus subsequently chose not to report their High School GPA, then the MAR assumption would be violated. Since the missing data were collected by the university *prior* to the collection of the other variables during the study, however, the MAR assumption was retained as plausible.

Under the MAR assumption, there are several commonly-applied strategies for handling missing data, including casewise deletion, pairwise deletion, substitution of means, regression predictions, full information maximum likelihood estimation, and multiple imputation (Collins et al., 2001). Though computationally less-intensive, the ad hoc techniques of casewise deletion, pairwise deletion, substitution of means, and regression predictions have “been shown conclusively to perform poorly except under very restrictive or special conditions” (Collins et al., p. 330). That is, simulation studies have demonstrated these techniques produce statistically biased estimates of variation within and between variables. Of the remaining two strategies, multiple imputation was chosen because of better average performance with small (N < 1000) data sets (Schafer & Olsen, 1998).

The multiple imputation strategy used for analyzing missing data in the study applied an iterative stochastic algorithm called Expectation-Maximization (EM). First developed by Dempster, Laird, and Rubin (1977), EM generates several imputed data sets based on the portion of the data set with complete data. In the technique used for this study (Gelman, Hill, Yajima, Su, & Pittau, 2009), the incomplete data set was “imputed” by replacing missing values by vectors of randomly adjusted means. Then, using the structural model as a base, each iteration applied two steps: (E-step) compute the expectation of the log-likelihood of the current estimated data set, (M-step) compute the parameters which maximize the log-likelihood from the E-step.

When the estimated data in an EM algorithm converges to within some small tolerance, the last completed data set is called an *imputed data set*. Imputed data sets are computed to replace missing data without introducing statistical bias into observed statistical power, variance, and associations among variables (Collins et al., 2001). Usually, 3-5 imputed data sets are constructed in this way (Schafer & Olsen, 1998), and subsequent statistical analyses are conducted separately on the imputed data sets. If results of statistical analyses are similar across imputed data sets, results are simply reported as the averages of results obtained from the separate analyses (Collins et al., 2001).

In the application of multiple imputation used in this study, the EM algorithm converged for three imputed data sets in 38 iterations. Each of the imputed data sets contained 60 (31%) imputes of missing ACT Math scores, 59 (30%) imputes of High School GPA, and 18 (9%) imputes of Required Math. Collectively, these imputes represented just 1.5% (137/8,970) of the entries in the data sets. The small proportion of imputed data led to nearly correlation structures that were identical to the hundred-thousandths place, so the results reported in forthcoming sections are means of results from the three imputed data sets.

Correlations between Indicators in the Structural Model

Following suggestions for the reporting of structural equation modeling (Bentler, 2007), this section reports on the correlation structure of the indicator variables in the study. As a standardized measure of joint variation which quantifies “closeness of linear relationship between two variables” (Snedecor & Cochran, 1989, p.177), the correlations are not meant to indicate the extent to which a variable “causes” or “predicts” another variable in the sense that might be inferred from experimental designs. Instead, since the model for directional effects between latent constructs was justified by the review of literature in Chapter III, correlations reported here simply quantify the extent to which observed indicators of the constructs are linearly associated.

Structural equation modeling was initially designed as a technique for analyzing the covariation structure of continuous indicator variables, but techniques have been developed to extend the structural modeling to all types of ordinal indicator variables

(Fox, 2006). That is, if one (or both) of two indicators is ordered but discrete, alternate estimates of correlation can be obtained under the assumption that the dichotomous and categorical variables reflect discrete levels of underlying continuous variables.

For example, it would not be possible to estimate correlation between gender (dichotomous) and ethnicity (categorical) because the fact that neither variable is ordered makes interpretation nonsensical. In contrast, the correlation between performance on a final exam item (dichotomous) and students' self-efficacy to complete the exam item (categorical) can be estimated by a polychoric correlation coefficient (Hair et al., 1998). Table 18 summarizes the types of correlations used in the analysis of indicator variables in the structural model.

Table 18.

Types of Correlation Used in the Structural Model

Type of Correlation	X × Y	Example
Product-moment	continuous × continuous	HS GPA × ACT Math
Biserial	dichotomous × continuous	Perf. Level 1 × HS GPA
Tetrachoric	dichotomous × dichotomous	Perf. Level 1 × Perf. Level 2
Polychoric	categorical × categorical dichotomous × categorical	SE Level 1 × Calib. Bias 1
Polyserial	categorical × continuous	SE Level 1 × ACT Math

The complete table of correlations between all 25 indicators is given in Appendix E. Probably due to the relatively large sample size (N = 195), many pairs of indicator variables were significantly correlated. In fact, 71% (214/300) of the possible correlations were significant at α = .01. Consequently, the magnitude and sign of the correlation coefficients were of primary concern. Table 19 summarizes the observed correlations between the composite measures of constructs.

Table 19.

Correlations between Indicators of High School Math Achievement, Math in Major, and Composites of Self-Efficacy, Calibration Bias, and Final Exam Performance

Measure	1	2	3	4	5	6	7
1. ACT Math	–	$.15^{**}$	$.31^{**}$	$.19^{**}$	$.28^{**}$	$-.32^{**}$	$.33^{**}$
2. HS GPA		–	$.52^{**}$	.05	$.18^{**}$	$-.22^{**}$	$.25^{**}$
3. HS Self			–	$.22^{**}$	$.23^{**}$	$-.25^{**}$	$.34^{**}$
4. Req. Math				–	.07	-.05	.04
5. Self-Efficacy					–	.10	$.39^{**}$
6. Calib. Bias						–	$-.45^{**}$
7. Final Perf.							–

Note. Self-efficacy, Calibration Bias, and Final Performance scales are each composite scales of seven indicators. HS GPA = high school grade point average (capped at 4.0); HS Self = self-assessment of high school math performance, Req. Math = semester math credits required by declared college major. $^{**}p < .01$.

As is common in this type of research (e.g., Pajares & Kranzler, 1995), magnitudes of correlations between indicators were typically between $|r| = .1$ and .5. The signs of significant correlations were, without exception, in line with expectations from the review of literature, including positive associations among indicators of high school mathematics performance, self-efficacy, Required Math, and final exam performance and negative associations between calibration bias and the other constructs. For example, increased calibration bias (tendency toward overconfidence) was associated with lower values on the high school mathematics performance indicators ($r = -.45$).

Correlation statistics are sensitive to departures from underlying assumptions of joint normal distributions, and are especially sensitive to outliers (Snedecor & Cochran, 1989). These bivariate normality assumptions were assessed graphically using scatter plots overlaid with contour ellipses corresponding to theoretical regions containing 30%, 60%, and 90% of data points under a normal joint-distribution with the observed correlation. Figure 15 shows a typical such scatter plot, indicating the moderate negative correlation (*r* = -.32) between ACT Math and calibration bias. The figure suggests the bivariate association does not differ substantially from assumptions of correlation. For example, 16 (8%) of the data points are outside of the 90% contour and 76 (39%) of the data points are outside of the 60% contour. No substantial violations of the bivariate normality assumption were identified in the analysis and (due to the restricted scales) no outliers were identified.

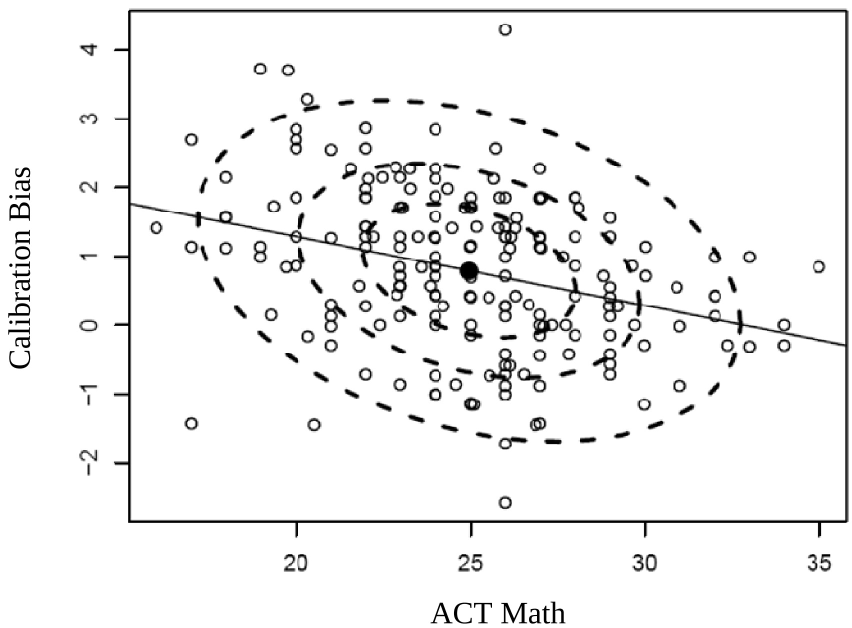

Figure 15. Scatter plot of composite Calibration Bias vs. ACT Math (*N* = 195). Plot shows contour ellipses corresponding to 30%, 60%, and 90% of data points under the assumptions of *r* = -.32 and the two variables have a normal joint-distribution.

Although Required Math was weakly correlated with ACT Math scores (r =.19) and self-evaluations of high school mathematics performance (r =.22), Required Math was not significantly correlated with *any* of the self-efficacy, calibration bias, or final exam performance indicators or composite scales. This, combined with the bimodal observed distribution, suggested the Required Math data were inappropriate for inclusion in the structural model.

Table 20 gives the correlations among the seven indicators of each of the self-efficacy, calibration bias, and final exam performance constructs. As in the composite measures, all correlations were positive. This joint variation among indicators of each construct indicated, for instance that a tendency toward reporting high self-efficacy for a survey item was moderately associated with increased self-efficacy to complete other survey items.

Table 20.

Correlations between Indicators of Self-Efficacy, Calibration, and Final Performance

Indicator	1	2	3	4	5	6	7
			Self-Efficacy				
1. SE Level 1	–	.38**	.35**	.37**	.28**	.16**	.29**
2. SE Level 2		–	.33**	.30**	.26**	.14**	.16**
3. SE Level 3			–	.35**	.36**	.39**	.38**
4. SE Level 4				–	.33**	.37**	.42**
5. SE Level 5					–	.38**	.34**
6. SE Level 6						–	.48**
7. SE Level 7							–
			Calibration Bias				
1. Bias Level 1	–	.05	.15**	.00	.04	.07	.08
2. Bias Level 2		–	.20**	.18**	.06	.07	.14**
3. Bias Level 3			–	.14**	.21**	.13**	.22**
4. Bias Level 4				–	.26**	.13**	.11*
5. Bias Level 5					–	.17**	.23**
6. Bias Level 6						–	.21**
7. Bias Level 7							–
			Final Exam Performance				
1. Perf. Level 1	–	.23**	.14**	.27**	.25**	.28**	.23**
2. Perf. Level 2		–	.32**	.47**	.40**	.21**	.50**
3. Perf. Level 3			–	.35**	.31**	.41**	.52**
4. Perf. Level 4				–	.29**	.35**	.61**
5. Perf. Level 5					–	.38**	.42**
6. Perf. Level 6						–	.49**
7. Perf. Level 7							–

Note. Heterogeneous correlations obtained according to variable type (see Table 21). $^{*}p < .05$, $^{**}p < .01$.

The correlations among indicators of calibration bias were relatively weak in comparison to the indicators of self-efficacy and final exam performance. For example, the mean inter-item correlation between indicators of calibration bias (r_{ave} = .14, *SD* = .07) was substantially lower than the mean inter-item correlations of the self-efficacy (r_{ave} = .32, *SD* = .09) and final exam performance (r_{ave} = .35, *SD* = .12) indicators. As mentioned in the reliability discussion, this may have been due in part to the cumulative effects of measurement error associated with the self-efficacy ratings and final exam performance scores whose differences determined the calibration bias scores. An alternate possibility is that the calibration indicators may have been considerably multidimensional in nature, so that, for example, Bias Level 1 would be better thought of as an indicator of a different construct than the construct associated with Bias Level 7. This possibility is considered in the following section on the analysis of the measurement model associated with the constructs high school mathematics, self-efficacy, calibration, and final exam performance.

Measurement Model

The hypothesized model assumed indicators of each of the five latent constructs satisfied a single-factor solution. Though the single-indicator specification of math in major made verification of the assumption of unidimensionality impossible (Hair et al., 1998), the analysis included evaluation of unidimensionality of the remaining latent constructs by fitting the measurement model (McDonald & Ho, 2002). A measurement model is the same as the structural model, except with paths between latent constructs (represented by ovals) omitted. Fitting the measurement model allowed for confirmatory common factor analyses of the assumptions to verify indicators of the four latent constucts were unidimensional.

Table 21 gives the standardized loadings of the indicator variables under the measurement model. Nearly all of the factor loadings were statistically significant, with the sole exception of Bias Level 1 (p =.12). Among the significant loadings, however, uniquenesses (the proportion of variation in the indicator variables unexplained by the single-factor models), were generally modest to high—ranging from .14 to .94 (*M* = .74, *SD* = .19). The proportions of common indicator variance explained by the one-factor models were significant in the models for High School Math (χ^2 (6, *N* = 195) = 24.7, $p <$ .001) and Self-Efficacy (χ^2 (14, *N* = 195) = 34.0, $p < .01$). However, the proportion of variance explained by the single-factor models were not significant in the cases of Calibration Bias (χ^2 (14, *N* = 195) = 11.4, $p = .62$) and Final Exam Performance (χ^2 (2, *N* = 195) = 7.6, $p = .91$).

Table 21.

Factor Loadings and Uniqueness for Single Factor Models of Self-Efficacy, Calibration Bias, and Final Exam Performance

Construct and Indicator	Standardized Factor Loading[a]	Uniqueness[b]	Proportion of Variance[c]
High School Math Achievement			.46**
HS Self	.99**	.01	
HS GPA	.50*	.73	
ACT Math	.29*	.90	
Self-Efficacy			.33**
SE Level 1	.51**	.74	
SE Level 2	.43**	.82	
SE Level 3	.63**	.61	
SE Level 4	.63**	.60	
SE Level 5	.56**	.68	
SE Level 6	.59**	.65	
SE Level 7	.63**	.60	
Calibration Bias			.15
Bias Level 1	.16	.98	
Bias Level 2	.29*	.92	
Bias Level 3	.46**	.79	
Bias Level 4	.38**	.86	
Bias Level 5	.48**	.77	
Bias Level 6	.35**	.88	
Bias Level 7	.47**	.78	
Final Exam Performance			.21
Perf. Level 1	.24**	.94	
Perf. Level 2	.44**	.80	
Perf. Level 3	.44**	.81	
Perf. Level 4	.54**	.71	
Perf. Level 5	.42**	.82	
Perf. Level 6	.44**	.81	
Perf. Level 7	.59**	.65	

Note. [a]Standardized factor loading = correlation between indicator and the latent factor. [b]Uniqueness = proportion of the indicator variance not explained by the latent factor. [c]Proportion of Variance = proportion of total variation in the indicators explained by the latent factor. $^{*}p < .05$, $^{**}p < .01$.

Alternate specifications of the models, including removal of indicators with the lowest standardized factor loadings and two- and three-factor solutions were fit to the data, but none of the alternative specifications explained a significant portion of the variance in calibration bias or final exam performance. Collectively, the factor analyses provided marginal evidence in support of the single factor assumptions for calibration and final exam performance, but also provided very little support for altering the structural model to incorporate additional latent constructs or sub-constructs.

Measurement models are commonly used to contextualize fit indices of structural models because fit indices from the baseline measurement model can be compared to fit indices associated with subsequent models which posit relationships among latent constructs (Fornell & Larcker, 1981). Though there are dozens of fit indices available for structural models that provide information regarding absolute fit, comparative fit, and parsimonious fit, Schreiber (2008) recommends reporting (1) overall chi-square, (2) comparative fit index (CFI), (3) the Tucker-Lewis non-normed fit index (NNFI), (4) standardized root mean square residual (SRMR), and (5) root mean square error of approximation (RMSEA).

CFI and NNFI are each standardized goodness-of-fit indices – values fall between 0 and 1 and generally indicate “good fit” if they exceed 0.9 (McDonald & Ho, 2002). In contrast, the overall chi-square, SRMR, and RMSEA statistics each measure the extent to which a model does not fit the data, so lower values of these indices suggest better model fit. The overall chi-square indicates divergence of the model from exact fit and is used primarily to compare nested models through likelihood ratio tests of $\Delta\chi^2$, or the change in overall chi-square (Schrieber, 2008). RMSEA values typically fall between 0 and 1; RMSEA below .05 indicates “good fit” and greater than 0.10 indicates “poor fit” (Schrieber). Finally, SRMR can be interpreted directly as the mean error of the model in reproducing correlations between indicators. For example, SRMR = .05 indicates the correlation matrix was reproduced to within about .05 on average (Schrieber, p. 828).

The overall measurement model chi-square was χ^2 (252) = 754.7, substantially lower than the chi-square of the independence (null) model (χ^2 (276) = 1492.8). Additional fit indices for the measurement model included CFI = .59, NNFI = .55, SRMR = .15, and RMSEA = .10. Though these fit values were obtained primary for comparative purposes, all the indices indicate an "inadequate" fitting model (Schreiber, 2008). That is, the measurement model, which assumes independence between latent constructs, is a poor model for the observed correlation structure of the indicator variables.

Specification of the Structural Model

An important obstacle to fitting the full hypothesized structural model was the inclusion of Required Math as an indicator of mathematics in major. As discussed in the earlier analysis of this indicator, (1) Required Math was severely non-normal with a bimodal shape and a large spike corresponding to mathematics majors, and (2) Required Math was not significantly correlated with any of the self-efficacy, calibration bias, and final exam performance scales. Inclusion of Required Math in the structural model led to consistent estimation of negative variances in the model, a practical impossibility referred to as Heywood cases (Hair et al., 1998). The Heywood cases persisted through attempts to transform the Required Math indicator to a normal distribution using arcsine and logistic transformations. Having failed two important assumptions of structural modeling – covariation with other indicators and normality – Required Math was omitted from the model specification along with its corresponding latent construct Math in Major.

After removing Required Math from the structural model specification, the estimation procedure for the restricted structural model based on the hypothesis converged in 210 iterations. All directional effects in the model were significant at the $\alpha = .05$ criterion with the exception of the posited direct effect of the latent variable High School Math Achievement on the latent variable Final Exam Performance ($\beta = -.19$, $p = .40$). Model fit indices included an overall chi-square of $\chi^2(246) = 608.0$, CFI = .70, NNFI = .67, SRMR = .08, and RMSEA = .09. The comparative fit indices (CFI and NNFI) were both below the .9 threshold for good fit, and the SRMR and RMSEA indices suggested marginal model fit. A likelihood-ratio test confirmed the structural model provided a significantly better fit than the measurement model ($\Delta\chi^2(6) = 146.7$, $p < .001$).

Several steps were taken to consider alternate model specifications, including model “trimming” to achieve improved parsimony and the inclusion of additional model paths. However, structural equation modeling is essentially a confirmatory statistical approach (Hair et al., 1998), so the analysis included a conservative approach to model re-specification. Inclusion of additional model paths was approached by inspection of Wald’s *W* statistics associated with modification indices (McDonald & Ho, 2002), but the largest modification indices were relatively small and were not theoretically supported. For example, the largest modification index was associated with estimation of correlated errors between Performance Level 6 and Bias Level 2 ($W = 16.8$), but the two indicators referred to entirely different exam items.

In contrast, there was some evidence to suggest removing some effects from the hypothesized model. For example, the measurement model suggested an insignificant loading of Bias Level 1 onto the Calibration Bias factor. However, the structural model which included the Bias Level 1 indicator found a slight but significant loading ($\beta = .21$, $p < .05$) and removal of the indicator did not produce improved model fit. Ultimately, the only specification changes retained in the final estimated model were (1) removal of Required Math and its associated latent construct Math in Major, and (2) removal of the non-significant path positing direct effects of High School Math Achievement on Final Exam Performance.

Estimation of the Structural Model

Figure 16 shows the final structural equation model with the estimated standardized directional effects; the estimates of measurement errors are omitted from the diagram for readability, but are presented along with the standardized coefficients in Table 22. Standardized parameter estimates for the structural model were all significant at the $\alpha = .05$ criterion, and values ranged from $\beta = .21$ (the loading of Bias Level 1 on Calibration Bias) to $\beta = .78$ (the loading of HS Self on HS Math Achievement).

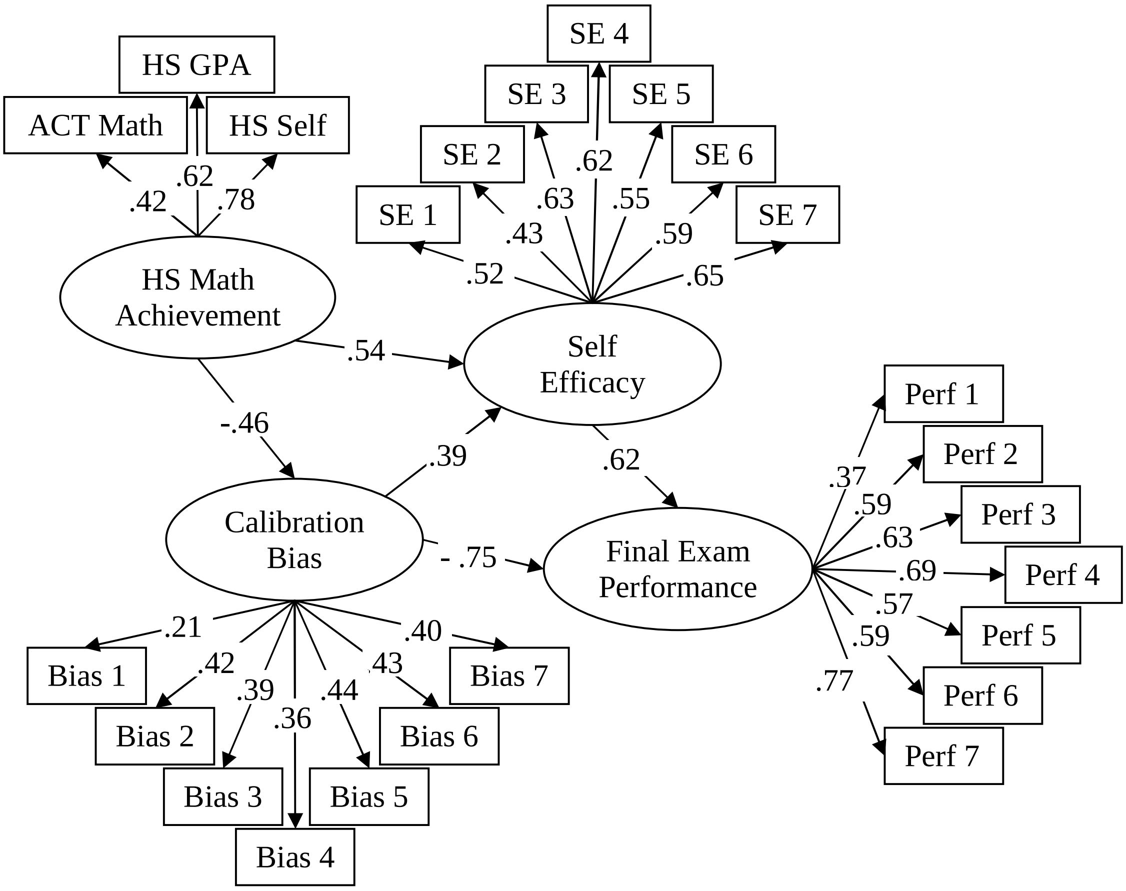

Figure 16. Standardized coefficients of directional effects in the final estimated structural equation model.

Table 22.

Standardized Parameter Estimates for Effects of Latent Constructs on Indicators

Construct and Indicator	Loading (β)[a]	Measurement Error[b]
High School Math Achievement		
HS Self	.78**	.38
HS GPA	.62**	.62
ACT Math	.42**	.83
Self-Efficacy		
SE Level 1	.52**	.73
SE Level 2	.43**	.81
SE Level 3	.63**	.60
SE Level 4	.62**	.61
SE Level 5	.55**	.70
SE Level 6	.59**	.66
SE Level 7	.65**	.58
Calibration Bias		.79
Bias Level 1	.21*	.96
Bias Level 2	.42**	.82
Bias Level 3	.39**	.85
Bias Level 4	.36**	.87
Bias Level 5	.44**	.80
Bias Level 6	.43**	.82
Bias Level 7	.40**	.84
Final Exam Performance		
Perf. Level 1	.37**	.86
Perf. Level 2	.59**	.65
Perf. Level 3	.63**	.61
Perf. Level 4	.69**	.52
Perf. Level 5	.57**	.68
Perf. Level 6	.59**	.66
Perf. Level 7	.77**	.40

Note. [a]β = estimated standardized effect of the latent factor on the indicator. [b]Measurement error = proportion of the indicator variance unexplained by combined latent effects. *$p < .05$, **$p < .01$.

Effects *between* latent constructs in the fitted structural model were interpreted as estimates of the sign and relative magnitude of effects posited by the model. For example, the review of literature supported direct effects of calibration bias on both final exam performance and self-efficacy, and the estimated coefficients suggested the negative effect of calibration bias (β = -.75) on final exam performance was comparatively larger than the positive effect of calibration bias on self-efficacy (β = .39).

Since the standardized path coefficients are multiplicative, the estimated indirect effect of calibration bias on final exam performance through its positive effect on self-efficacy was β = .39(.62) = .24, meaning the large direct negative effect of calibration bias on final exam performance was mediated somewhat by the indirect positive effect coming from the effect of bias on self-efficacy. Similarly, though the modeling did not identify a direct effect of high school math achievement on final exam performance, high school math achievement did have indirect effects on final exam performance through the separate effects of high school mathematics achievement on self-efficacy and calibration bias. Table 23 summarizes the direct, indirect, and total effects identified in the fitted structural model.

Table 23.

Standardized Direct, Indirect, and Total Effects between Latent Constructs

Effect of…	on…	Direct	Indirect	Total
HS Math Achievement				
	Calibration Bias	-.46**		-.46
	Self-Efficacy	.54**	-.18	.36
	Final Exam Performance		.57	.57
Calibration Bias				
	Self-Efficacy	.39*		.39
	Final Exam Performance	-.75**	.24	-.51
Self-Efficacy				
	Final Exam Performance	.62**		.62

Note. $^{*}p < .05$, $^{**}p < .01$.

Though the estimated effects between latent constructs were of primary concern, the coefficients from latent constructs (ovals) to indicator variables (rectangles) given in Figure 16 and Table 23 can be used to develop a qualitative understanding of the constructs labeled by high school mathematics achievement, calibration bias, self-efficacy, and final exam performance. For example, the descending order of effects of high school math achievement on High School Self (β = .78), High School GPA (β = .62), and ACT Math (β = .42) suggested that the latent construct was a mixture of cognitive commonalities among the three indicators, with perhaps greater focus on participants' self-perception of their performance in high school mathematics classes. Similary, the final exam performance construct could be considered more related to students' performance on the "easiest" sampled final exam item (Perf. Level 7, β = .77) than the students' performance on the "hardest" sampled final exam item (Perf. Level 1, β = .37). Indicators loaded onto the self-efficacy and calibration bias constructs approximately equally, though the comparatively smaller loadings of calibration bias indicators were further evidence of relatively larger uniqueness components among the calibration indicators.

The structural model contained three endogenous latent constructs – calibration bias, self-efficacy, and final exam performance – and the model fit included estimates of the proportion of unexplained variation in each of the constructs. The model accounted for an estimated 21% of the variation in calibration bias and 25% of the variation in self-efficacy. This relatively high proportion of unexplained variation in self-efficacy and calibration bias reflects the relatively few exogenous variables in the model. For example, the only construct posited to have an effect on calibration bias was high school mathematics achievement, which included only three indicators. In contrast, the model accounted for 83% of the variation in the final exam performance construct.

As in the measurement model and initial structural model specification, the final structural model results should be considered in light of indices of model fit. The overall model chi-square of 609.0 on 247 degrees of freedom indicated the model differed significantly from the exact (saturated) solution and comparative fit indices (CFI = .70, NNFI = .67) indicated less than adequate model fit. However, a likelihood-ratio test confirmed the structural model provided a significantly better fit than the measurement model ($\Delta\chi^2(6) = 146.7$, $p < .001$) and the observed RMSEA value (0.086) fell between the thresholds for good fit and acceptable fit. The most easily interpreted index of model residuals, SRMR = 0.075, indicated the fitted model reproduced correlations among indicator variables to within an average of .08. Of the 276 correlations in the final structural model, correlations ranged from -.39 to .61 ($r_{\text{ave}} = .09$, $SD = .19$), so the average error of .08 in the predicted correlations was considered marginally acceptable. In summary, the model fit indices indicated the structural model explained a substantial portion of the correlation structure of the indicator variables, but there was also substantial unexplained variation in the data that may lower the statistical power of findings.

The structural modeling results suggest partial answers to the following four research questions and corresponding hypotheses.

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects of participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

H1 High school mathematics achievement will have a moderate positive effect on the amount of mathematics in participants' college major.

H2 Both high school mathematics achievement and the amount of mathematics in participants' college major have small positive effects on participants' calibration.

H3 High school mathematics achievement and the amount of mathematics in participants' college major will have moderate positive effects on self-efficacy. Calibration will have a small negative effect on self-efficacy.

H4 High school mathematics achievement and the amount of mathematics in participants' college major will have small positive effects on mathematics performance. Calibration will have a large negative effect on mathematics performance. Self-efficacy will have a large positive effect on mathematics performance.

As in the structural modeling, the results are mixed regarding the hypothesized effects associated with math in major. The analysis of correlations among indicator variables provided some evidence that students' high school mathematics achievement may have a small positive effect on the amount of mathematics in students' declared college major; Required Math was weakly correlated with both students ACT Math scores (r =.19, $p < .01$) and students' self-assessments of their performance in high school mathematics classes (r =.22, $p < .01$). In contrast, correlations suggested limited evidence in support of associations between Required Math and calibration, self-efficacy, or final exam performance indicators.

The limited associations between Required Math and the other indicator variables, together with distributional characteristics which made Required Math poorly-suited for structural modeling, led to the removal of Required Math and its associated construct, math in major, from the structural model. While suggesting no effects of math in major on the other constructs (Q2), the removal of math in major from the structural model necessarily resulted in inconclusive findings regarding possible effects of high school mathematics achievement on math in major (Q1).

The table of estimated direct and indirect effects (Table 23) from the structural model provides much of the evidence regarding the hypothesized effects among high school mathematics achievement, calibration bias, self-efficacy, and final exam performance. High school mathematics achievement had no direct effect on final exam performance, but evidenced approximately equal and opposite moderate effects on calibration bias ($\beta = -.46$) and self-efficacy ($\beta = .54$). Accounting for both direct and indirect effects, high school mathematics achievement had a moderate negative effect on calibration bias ($\beta = -.46$), a slightly smaller positive effect on self-efficacy ($\beta = .36$), and a moderate positive effect on final exam performance ($\beta = .57$). These findings support the direction and significance of hypothesized effects, but differ to some extent from the expected magnitudes. For example, the hypothesized moderate effect of high school mathematics achievement on self-efficacy was decreased by the indirect effect mediated by the relationship between high school mathematics achievement and calibration bias.

Calibration bias had a relatively small positive direct effect on self-efficacy ($\beta = .39$) and a relatively large direct negative effect on final exam performance ($\beta = -.75$). However, including the indirect effect of calibration bias on final exam performance through self-efficacy, the total effect of calibration bias on final exam performance was moderately negative ($\beta = -.51$). Interestingly, the observed positive effect on self-efficacy was the opposite of the hypothesized relationship, suggesting that less-calibrated students tended to be *more* confident in their abilities to complete the sampled final exam items correctly. Finally, self-efficacy had a relatively large positive effect on final exam performance ($\beta = .62$). This substantiated the hypothesized effect of self-efficacy on exam performance.

Themes from Qualitative Interviews

The concurrent mixed methods research design included task-based interviews of 10 prospective secondary mathematics teachers. These interviews, constructed with the help of the participants' instructors, included mathematical tasks similar to midterm exam items and discussion of students' understandings of the related mathematical concepts. The interviews focused on the participants' (1) reported self-efficacy to complete tasks in their mathematics classes correctly, (2) reasoning for choosing self-efficacy ratings, and (3) experiences in prior college mathematics classes which may have affected the participants' self-efficacy to complete university mathematics. Special emphasis was placed on the variety of mathematical competencies, self-efficacy, and college mathematics experiences observed across the interviews. Building on the description of interview participants provided in the methodology chapter, the following narrative includes descriptions of the themes which emerged from the qualitative data analysis along with vignettes and quotations that illustrate and support the qualitative claims.

Overview of Themes from Task-Based Interviews

The focus of the thematic coding and synthesis of task-based interview data was on developing an understanding of the variety of secondary mathematics majors' experiences to inform interpretations of the structural modeling results. Several themes emerged from analysis of the task-based interviews, and the themes were framed in the conceptual framework of social cognitive theory and the review of literature. Alignment between the themes and the structural model supported triangulation of findings between the quantitative and qualitative data sources and cross-validation of the measures.

Five qualitative themes were identified from the task-based interviews of 10 secondary mathematics majors. These included (1) strong high school mathematics performance, (2) lowered self-efficacy associated with perceived low exam performance, (3) content-specific evaluations of self-efficacy for interview tasks, (4) tendency toward slight overconfidence with improved calibration on low self-efficacy items, and (5) increased self-efficacy to complete a mathematics course after initially not passing a course.

Theme #1: Strong High School Mathematics Performance

The interview participants generally reported successful experiences in their high school mathematics classes. All but one of the participants enrolled in mathematics during all four years of high school. The exception, Megan, took Advanced Placement Calculus as a high school junior, but chose not to take mathematics during her senior year. Justin began taking high school mathematics classes in middle school, and all of the interview participants completed at least pre-calculus mathematics in high school. In fact, 8 of the 10 participants (all except Heather and Matthew) completed mathematics classes in high school that included opportunities for college credit. Four participants completed college algebra, four completed calculus, and college statistics and trigonometry were each completed by one student. Sarah and Jackie both completed two college-level mathematics classes in high school.

Justin's positive self-evaluation of his mathematics performance and self-efficacy in high school was typical of the participants: "I thought I was pretty good at math in high school. On a scale of 1 to 10, I'd give it about an 8. With some time, I felt like I could figure things out." In particular, several of the interview participants described perceived benefits from high school mathematics preparation in subsequent college mathematics classes. Jennifer believed her success in Linear Algebra was due in part to her high school mathematics preparation, saying, "I thought Linear Algebra was kind of easy. I think because I did a lot of algebra in high school. Like the matrices – I did a lot of that in high school." Besides experience with college-level content, some students (Emily, Jackie, and Heather) pointed to study habits and problem-solving skills developed in high school mathematics as sources for confidence in their college mathematics coursework.

The four students who completed calculus in high school all chose to begin with first-semester calculus in college, and all cited their high school calculus experience as beneficial to their performance in college mathematics. Elizabeth drew on her high school calculus experience when she enrolled in Business Calculus as a freshman, a choice that ultimately led to her decision to become a high school mathematics teacher:

> I had taken calculus in high school, so I felt pretty confident and I was actually tutoring some of the seniors in my Business Calculus class. I really liked that class. It wasn't just that I was learning, but I actually wanted to do my homework. I'd usually do it like an hour after class. That's the only class where that's happened. That's the class that made me want to be a secondary math major.

There were few exceptions to the general theme of strong performance in high school mathematics among the interview participants. Nicole was the only participant who failed a mathematics class prior to college, saying "I did fine [in high school math], but I started to slack because I was a teenager, and I thought school didn't matter, and I had to take Pre-Calc twice. I understood it and I would do the homework, but I wouldn't turn it in, and my grade just dropped." Besides Nicole's experience in Pre-Calculus, the only other participant who described any kind of low performance in high school mathematics was Elizabeth, who said she experienced some difficulties in college calculus because she was "never really good at trig in high school."

Theme #2: Effects of Perceived Low Exam Performance on Self-Efficacy

The interview transcripts were coded for instances of Bandura's (1997) four sources of self-efficacy, which include mastery experiences, social persuasions, vicarious experiences, and physical and emotional states. Coded excerpts supported each of the four sources, but the participants' descriptions of their college mathematics self-efficacy supported mastery experiences and vicarious experiences as the primary sources of mathematics self-efficacy. In particular, perceived exam performance – both personal exam scores (mastery experiences) and the perceived performance of peers (vicarious experiences) – appeared to have primary effects on participants' mathematics self-efficacy. However, participants' personal feelings about their instructors, especially perceptions of approachability, appeared to mediate the degree to which low exam performance affected mathematics self-efficacy.

The variety of students' interpretations of their exam performance in college mathematics classes is exemplified in the descriptions of Calculus III offered by Sarah, Emily, Jennifer, and Elizabeth. The four students were all enrolled in the same section of Calculus III about a year prior to the interviews. Jennifer, Emily and Elizabeth passed the class with grades of B, B, and C, respectively, and Sarah earned an F. Despite the variety of grades, the students described very similar feelings of surprise and disappointment in their exam scores, and their perceptions of that exam performance seemed to have qualitatively different effects on their mathematics self-efficacy.

For Jennifer, low perceived performance on Calculus III exams appeared to have little lasting effects on her mathematics self-efficacy. Jennifer recalled feeling encouraged to become a math major because of high exam scores in classes like College Algebra and Trigonometry, saying “I think [those classes] helped me decide, ‘Hey, I’m good at math. Like, I’m better than most of the students in my class.’” Having earned an A in four college mathematics classes prior to Calculus III, Jennifer expressed surprise when she scored below 70% on 3 of the 4 exams in Calculus III. Though she earned a B in the class, Jennifer struggled to explain her unexpectedly low exam performance, ultimately attributing her lack of understanding to the instructor’s teaching style:

> I didn’t hardly understand Calc III. I don’t know what it was. I tried and everything, and I’m a good studier… It could have been the professor’s teaching style. You know how it is with math, it kind of depends on the teacher who’s teaching it, how well you do. I kind of would say that was the main thing.

Sarah’s description of her performance in Calculus III was even more closely tied to her personal feelings about the teacher than Jennifer’s description of the class. Like Jennifer, Sarah had grown accustomed to high performance in college mathematics classes prior to Calculus III and had earned an A in both Calculus I and Calculus II. Asked about her exam performance in Calculus III, Sarah said all her exam scores were below 70% and explained the low performance almost exclusively in terms of feelings about the instructor:

> I understood Calculus III, but I really don’t like the instructor. So, at the end of the semester, I was just like I am willing to just shoot myself in the foot to not do any more work for you. And that was basically what I did. I really, I was not happy… Actually, my Calc III teacher was the only math teacher I’ve had that I didn’t like.

Sarah took Discrete Math and Calculus III during the same semester, and she perceived her exam scores to be low in both classes, but the exams in the two classes seemed to have disparate effects on her mathematics self-efficacy. When Sarah scored about 50% on her first Discrete Math exam, she found it difficult to understand the letter grade of B posted next to what she thought was as a failing score. She said, "[The Discrete Math Instructor] told us it was normal that everyone failed, but it made me feel really bad about it. I did decently, but I felt like I was doing really, really, bad." Despite this disappointment, Sarah said she understood the grading system and she found the instructor approachable, "[The instructor] was really hard, but I felt like I could talk to him." In contrast, Sarah did not feel comfortable talking to the Calculus III instructor, and the first exam was much more disappointing:

> I think our first test in Calc III, I think like six people passed it out of both classes. [The instructor] didn't say he would curve or anything. He was like, I'll drop the lowest test grade, but it just kind of puts in your mind, this is how you're going to do on all the tests. I was one of the people who failed it, like everyone else.

Sarah's perceptions of her exam performance and the vicarious experience of similarly low perceived performance of her peers seemed to convince her that she would continue to perform poorly in Calculus III. At the same time, she described a continuing calibration bias toward overconfidence on the exams in Calculus III. Prior to the first exam, she believed she could probably earn a B, but thought the test was unfair: "[the instructor] threw in a lot of tricks and things from way back when, like 8th grade algebra." By the end of the semester, Sarah said she decided to give-up on trying to pass the class:

> In Calc III, toward the end, my grades were not improving, and I felt like I knew [the material]. I don't know why I didn't know it, because I should have been able to do it… Then there were only a few assignments left, so I was like, the highest grade I can get is a C minus, so there's no reason to stress about it. I was just like,

> I’ll take an F and not worry about it. I still went to class, I still did the homework. Our last lab, I was like “eh.” And the final, I didn’t even study for it. I knew I couldn’t get anything higher, so I just went and took it.

Sarah’s low exam performance and high calibration bias on exams seemed to contribute to self-handicapping behavior and a failing grade in Calculus III. The low performance also resulted in social consequences, including questions from family members about her choice to become a mathematics teacher. As Sarah described it,

> Calc III was definitely a downer. I have a lot of people in my family who make fun of me for that. I was like, “Yeah, I’m not doing really well in Calc III.” And they’re like, “And you’re going to be a math teacher?” and I just say, “Well, the good thing is I’m never going to be teaching anything that high.”

Emily, a junior who decided to become a mathematics teacher at the age of 16, experienced similar disappointment in her Calculus III exam scores. When talking about Calculus III, she remembered questioning the choice to become a mathematics major:

> I had problems. Failed the tests. It was horrible. That was when I was like, "I don't know if I should stay a math major. If I can't understand this, I'm going to get into higher math, and it's just not going to click.” It just scared me. That was not a good semester. Those were not happy nights.

Though Emily earned a B in the Calculus III class, she said the class was a turning point for her exam performance. In four classes she had completed since Calculus III, Emily had come to expect exam scores between 50% and 80%: “One thing I'm learning to accept right now is I don't do as well on tests as I used to.”

Emily described her experience in Calculus III as persistence through confusion and disappointment. She described taking notes during lecture, reading the textbook, completing all the homework exercises in the book, and working with classmates on the study guides provided by the instructor for exams. She found it difficult to connect the drawings and equations from her notes with the material in the textbook, but thought she understood the homework and study guides. When she got to exams, however, Emily described disappointment and surprise:

> The first test, well over half the people failed that test, and I think the next test as well. I think the highest test score I got in Calc III was a C. The lowest was a 39%. It was so bad. But somehow I got a B. I don't know how. [The instructor] might have curved the tests, but I don't know. Homework I did well on, because when the book presented it, I could get it. The problems on the test, though, I just couldn't see a relationship between them… So, I had problems. Failed the tests.

Though they earned very different final grades in the class, Emily and Sarah described some common mechanisms that influenced their self-efficacy, including (1) vicarious experiences as a source for lowered self-efficacy on exams, (2) calibration bias in the form of overconfidence in the ability to perform well on exams, (3) calibration bias in the form of underconfidence to earn the grade they wanted from the class, (4) emotions associated with the fear of being able to do higher-level mathematics, and (5) lowered self-efficacy to complete a mathematics major.

While Jennifer and Sarah focused on their dislike of the teacher or teaching style, Emily's description of her low performance also included her strategies for overcoming the low exam scores she was experiencing. Emily developed several new strategies to improve her performance in Calculus III, including (1) learning to use the textbook when she could not understand the instructor, (2) working with a study group on homework and study guides, and (3) asking questions in class when she knew she was not the only person who was confused. She also described a lasting change in how she viewed grades, saying "I think I could get a C and be proud of it, if I know that I worked hard enough. Not to blame the professor, but if the class is with a professor that I didn't learn well with, but I still know that I tried, then I'd be happy with a C." While Emily's exam grades in Calculus III did not improve during the semester, she found benefits in the studying strategies she developed and reported using the strategies in subsequent mathematics classes.

Yet another view of the role of exam performance in mathematics self-efficacy can be found in Elizabeth's experience of Calculus III. Elizabeth initially experienced success in the class, earning 98% on the first exam. She attributed her early success to taking Calculus II from the same instructor as well as to almost daily meetings with a study group – later joined by Emily – which carried-over from the Calculus II class. Nonetheless, Elizabeth said she earned a C in Calculus II and disliked the instructor from that prior experience, so she expected to earn a C in Calculus III. Elizabeth earned "really low" scores on the remaining exams in the class, and she remembered inconsistent performance and attendance. Elizabeth relied on the study group meetings to complete homework and learn the content, and she described a gradual decline in her self-efficacy to learn new content in the class, "At the beginning I felt like I could learn the math, but by the end it was just overwhelming."

As suggested by the four participants' experiences in Calculus III, students' descriptions of the trajectory of their mathematics self-efficacy through college mathematics classes pointed to low exam performance and comparisons to perceived performance of peers as primary sources of mathematics self-efficacy. Interview participants also described a close link between perceptions of exam performance and their personal feelings about teachers. In particular, the participants said they had more positive views of low exam performance when they liked the instructor on a personal level or felt the instructor was approachable or friendly. Emily described feeling encouraged to work past her confusion in Discrete Math because she liked the instructor and thought he was personable and interested in the students, and Sarah described a willingness to accept low performance in Probability because she liked the instructor: "I like my Probability teacher, she's nice. It's a hard class, but I can still talk to her if I have questions. It makes me feel better about my C that she's actually nice."

Elizabeth's description of her Geometry instructor mirrored Sarah's view on Probability:

> Right now, I'm in Geometry. I don't know anything about the instructor, but I really like her. Maybe it's just that she's more friendly. I don't know. I find that the teachers I care about as people, I also care about what they have to say. The teachers I don't think highly of, I really don't want to listen to them in class. And it's not because, well, this person gives too much homework, so I don't like them. Because, like, my Geometry teacher, her homework is really intense, but I really like her. But my Calc III teacher, his homework was intense too, but I don't like him. I think it's the personality.

Theme #3: Content-Specific Reasoning for Self-Efficacy Judgments

Each of the task-based interviews included a self-efficacy survey similar to those used in the quantitative strand of the study. Participants completed the surveys, which included self-efficacy ratings of 7 to 11 tasks developed in conjunction with instructors to be similar in difficulty and content to exam items. (See Table 24 for composite ratings.) While there was some evidence of response styles among the participants, students' explanations of their self-efficacy ratings generally supported the validity of the ratings as representing content-specific self-efficacy beliefs.

The participants typically described the reasoning for their self-efficacy ratings in very content-specific terms, referring to prior experience with the tasks, anticipations of the number of steps required to complete the tasks, or familiarity with the content. For example, Megan and Heather both described a tendency to give lower ratings on calculus problems that involved trigonometric functions because of past difficulties differentiating functions that contained sine or cosine. Jennifer rated one item as a “2” out of 6 because she recalled not being able to solve a similar problem on a recent exam. Referring to a Calculus II problem asking for the volume of a solid of revolution formed by rotating a region bound by a parabola about the *y* axis, Jackie said, “I would put a 4 because I don’t know what the question is asking, but I think I understand it. So, I’d graph it, and try to see what they’re asking.” Each of these patterns of reasoning is consistent with the social cognitive view of self-efficacy as a content-specific assessment of one’s ability to complete a performance task (Bandura, 1997).

There was evidence that some participants had aversions to responding with the highest rating (6) or lowest rating (1) listed on the self-efficacy surveys. For instance, none of the respondents rated their self-efficacy on the survey tasks with the lowest available rating. Heather’s explanation of her reticent to choose the lowest rating reflected the common response that there is always a chance of solving an exam problem: “To put a 1, you’d have to put something I’ve completely never seen before, for me to believe there’s no possible way for me to get it. As long as I’ve seen that kind of math before, I figure there’s at least a possibility I can get it.” Consequently, Heather’s ratings were effectively limited to the range of 2 to 6.

Several respondents (Jennifer, Megan, Emily, Sarah, and Nicole) described a belief that they would rarely, if ever, rate an exam problem in their class with the highest possible rating (6). Sarah summarized her reasoning for not using the highest self-efficacy rating as reflecting a general belief that there is always a chance of making a mistake in a mathematics problem:

> I just usually don’t feel that way during a test. I really don’t. I’m always like, I can probably miss a couple points on that. Even if I do get it completely right. I

> got a few of the problems right on my last test, but I wouldn't put a 6 next to them.

Emily said her reticence to report the highest self-efficacy was linked to a general mistrust of her feelings of confidence. Asked why she did not rate any of the 10 problems on the self-efficacy survey with a 6, Emily said:

> That's me. That's just how I always am. I do have a problem with trusting myself. Even if I know I'm doing it right, there's always something in me saying... I guess it's kind of like trusting your instincts. I'm just not good at that. It's why I have problems with multiple choice, because I just don't trust myself. If this problem were on a test, though, I'd leave it at that. I'd move on.

Theme #4: Calibration Bias toward Slight Overconfidence on Interview Tasks

To explore calibration bias in the interview setting, participants in each of the task-based interviews completed at least two tasks selected from those on their self-efficacy survey. In particular, the participants chose at least one task to complete from among the survey items in which the participant provided low self-efficacy ratings (1 or 2) and at least one task from among the tasks rated with high self-efficacy ratings (5 or 6). Table 24 outlines the performance of the interview participants on the sampled tasks along with qualitative descriptors of the participants' observed calibration and mean self-efficacy ratings on the survey items. Collectively, participants seemed to be more calibrated on items for which they expressed low self-efficacy. Nine of the 10 participants incorrectly solved the problems for which they expressed low-efficacy, while only 5 of the 10 participants correctly solved the problems for which they expressed high self-efficacy.

Table 24.

Performance of Interview Participants on Tasks with Low and High Self-Efficacy Ratings

		Performance Tasks		
Participant	Mean SE[a]	Low SE	High SE	Calibration
Heather	3.9	0	0	moderately overconfident
Matthew	3.6	0	0	moderately overconfident
Megan	2.9	0	0	slightly overconfident
Justin	3.6	0	5	calibrated
Jackie	3.9	0	5	slightly overconfident
Nicole	2.3	0	0	calibrated
Sarah	2.7	0	0	slightly overconfident
Jennifer	3.4	0	5	calibrated
Emily	2.9	5	5	moderately underconfident
Elizabeth	3.1	0	5	calibrated

Note. [a] Mean SE ratings are on a scale of 0 to 5 and reflect 7 to 11 ratings. Calibration descriptors are holistic qualitative assessments. SE = Self-Efficacy. On performance tasks, 0 = incorrect, 5 = correct.

The qualitative assessments of calibration presented in Table 24 were based on the students' performance on interview tasks, self-efficacy ratings, and their descriptions of the reasoning for self-efficacy ratings. Four of the participants appeared to be well-calibrated, one participant (Emily) demonstrated moderate underconfidence, three participants showed slight overconfidence, and two students demonstrated moderate overconfidence.

Since every reported self-efficacy rating during the interviews was above the lowest available value (2 or above) and performance was scored on a dichotomous (correct/incorrect) scale, any incorrect attempt on a performance task would numerically corresponded to a positive calibration bias score, which was operationally defined as overconfidence. To understand the qualitative meaning of the numerical calibration scores, the inquiry included analysis of conceptual understandings and procedural skills that contributed to incorrect attempts. Nearly all of the attempts recorded as incorrect were identified as inability to set-up the solution (5 of 14), misinterpretation of the task requirements (5 of 14), or inaccurate application of a procedure (3 of 14). Only 1 of the 14 incorrect attempts was the result of an arithmetic or algebraic error.

Megan's attempt to sketch the graph of a function from a graph of the derivative of the function was a typical example of performance which indicated potential calibration bias. Megan chose the task as an example of a problem she felt very confident to complete correctly and provided her highest self-efficacy rating (5 out of 6) on the task. When asked to complete the task, however, Megan applied a procedure to graph the derivative of a function from the graph of the function (the inverse procedure). Megan quickly applied the incorrect procedure accurately, successfully producing an approximate graph for the second derivative of the function. When the researcher explained this error, Megan was able, with some help regarding the role of maxima and minima, to complete the initial task correctly. This performance, together with Megan's self-efficacy ratings on the other tasks and explanations about her reasoning for selecting self-efficacy ratings, contributed to the choice of the qualitative descriptor "slightly overconfident" for her calibration in Calculus I at the time of the interview.

Interestingly, the four students enrolled in Probability collectively demonstrated very little calibration bias during the interviews and the three students enrolled in

Calculus I each appeared to be overconfident in their assessments of self-efficacy. Some plausible sources for this observation include (1) instructional differences between the two classes, (2) improved calibration as a result of additional college mathematics experience, (3) differences in the relative difficulty of the self-efficacy items, and (4) chance (due to the small sample).

Theme #5: Effects of Failing College Mathematics Classes on Self-Efficacy

Table 25 summarizes the enrollment history of the interview participants in mathematics classes since beginning their undergraduate education. Of the 10 interviews, two students (Heather and Matthew) were enrolled in their first college mathematics class at the time of the interview. Of the remaining eight students, five had failed at least one mathematics class in college, accounting for a total of nine failed college mathematics classes.

Table 25.

Mathematics Enrollment History of Task-Based Interview Participants

Participant	Calculus I	Calculus II	Calculus III	Discrete Math	Linear Algebra.	Probab-ility
Heather	In					
Matthew	In					
Megan	Fail, In					
Justin	Fail, Pass	Fail, In		Fail, Pass		Fail, In
Jackie	Pass	In		Pass		
Nicole	Fail, Fail, Pass	In				
Sarah	Pass	Pass	Fail, Pass	Pass	In	In
Jennifer	Pass	Pass	Pass	Pass	In	In
Emily	Pass	Pass	Pass	Pass	Pass	In
Elizabeth	Pass	Pass	Pass	Fail, Pass	Pass	In

Note. In addition, Megan, Jennifer, and Jackie had each passed College Algebra and Sarah, Elizabeth, and Emily were each enrolled in Geometry at the time of the study; Jennifer had passed Trigonometry and Geometry. In = enrolled in the class at the time of the study; Fail = earned grade of D or F in a prior semester; Pass = earned grade of A, B, or C in a prior semester.

When asked to describe the classes in which they did not pass, participants referred to (1) dislike of the instructor or teaching style, (2) perceived personal learning styles, and (3) insufficient preparation in prior mathematics courses. Of particular interest were the participants' perceptions of these experiences on their self-efficacy to succeed in college mathematics.

In contrast to the pattern of lowered self-efficacy following low exam performance, the participants typically described increased self-efficacy after not passing college mathematics classes. Elizabeth, a junior who earned an F in Discrete Math during her sophomore year, said she performed much better in Discrete Math the second time because of increased familiarity with the content and a perception that the second instructor was more approachable: “What I had struggled with the first time seemed so much easier with the new instructor. I think it was easier because I knew what was coming a little bit more, but I also think it was because the teacher was less scary.”

Nicole, who failed to pass Calculus I twice prior to passing the class during her sophomore year, was explicit about the benefits she perceived from her history of difficulties in calculus. Nicole attributed her challenges to poor algebra skills dating back to an ineffective eighth grade algebra teacher, but she emphasized what she learned about herself as a student. In particular, she had learned the shortened format of a summer mathematics class was not useful and that she needed to talk to instructors when she got confused. She said the calculus was easier the third time because of the previous “multiple go-rounds” and because she liked the instructor. Overall, she expected the struggles in Calculus I would help her as a future middle school mathematics teacher:

> I figure, I have all this struggling history, so if I can make it, I can relate to the students more… They will be like, “I’m sure you were always good at math”, and I’ll be like, “Actually, I had to take Calculus 3 times.” I took it my freshman year, I took it during the summer, and I took it last semester. So, I wasn’t really that good, and I did it.

Among the participants, Justin, a junior enrolled in Calculus II, had the most positive outlook on not passing college mathematics classes. Prior to the interview, Justin had earned an F in every college mathematics class the first time he enrolled, including Calculus I, Calculus II, Discrete Math, and Probability. He described “checking-out” of Calculus II when there seemed like there was too much work, feeling unprepared for Probability, and falling behind after missing classes in Discrete Math and Calculus I1.

Referring to his second attempt at Calculus II, Justin said, "When I miss class, it causes a lot of problems. Like, a couple weeks ago I just wanted to go hang out with my friends. So, I left on a Monday, and by the time I came back the next week, I'd missed Chapter 7."

Despite his history of poor performance, and the year in which he took no mathematics classes while he was on academic probation at the university, Justin expressed very high mathematics self-efficacy. He had been advanced in mathematics coursework since elementary school, had taken high school mathematics classes while in middle school, and rated his self-efficacy in mathematics upon graduating high school as 8 out of 10 (on a scale of 1 to 10). Justin rated his overall self-efficacy to learn a new mathematics topic as a 10 out of 10 and rated his self-efficacy on eight tasks in Calculus II with ratings between 3 and 6 (out of 6).

In second attempts at Calculus I and Discrete Math, Justin passed both classes with a letter grade of C. He said he was "in no hurry" to graduate and that he planned to continue taking mathematics classes more than once as needed. He described his low grades in college mathematics classes as being primarily the result of a personal learning style that benefited from taking classes more than once:

> When I'm learning math, I feel like it takes me just a little time to start absorbing the information. I feel like, now [in Calculus II], it's starting to sink in… it just takes me a couple times. I guess I need to preview the class before I can get it. It's not that I can't learn it, it's just that I need to be shown what I'm doing first.

The "preview effect" described by Justin, Nicole, and Elizabeth appeared to increase the participants' self-efficacy to complete the mathematics classes in subsequent attempts. Sarah, whose experience earning in F in Calculus III is described in the theme regarding exam performance, also expressed a positive view of the familiarity she gained with the Calculus III content when she described her second attempt at Calculus III. Though Sarah, Nicole, and Elizabeth each expressed disappointment in their initial performance, their prior experience appeared to leave them (as well as Justin) with increased self-efficacy to pass the classes in subsequent attempts. Though the possibility of failing to pass a college mathematics class having the effect of lowering students' self-efficacy in other mathematics classes seemed plausible from a social cognitive theory perspective, there was limited evidence in the interview data to support specific negative effects of not passing college mathematics classes on the participants' mathematics self-efficacy.

Though the quantitative results and qualitative themes have been presented in this chapter separately, the secondary mathematics majors in the qualitative interviews can be viewed as an important subset of the participants in the quantitative strand of the inquiry. In addition, the use of a common conceptual framework in the quantitative strands, together with the contextual data collected in both strands of the inquiry, allowed for convergence of the quantitative and qualitative findings. In the next chapter, the qualitative themes and quantitative findings are synthesized, compared, and contrasted. In addition, the discussion of results includes further discussion of limitations in the study, connections to related literature, and potential implications of the study.

CHAPTER V

DISCUSSION

The social cognitive approach to learning rests on the core idea that "people are at least partial architects of their own destiny" (Bandura, 1997, p.8) in the sense that they work to shape their environment based on perceived opportunities for success. Self-beliefs are central to this view, because self-efficacy and related beliefs act in concert with knowledge and competencies to affect academic performance (Pajares & Urdan, 2006). A large body of self-efficacy research has established its prominent role in academic behavior: "Students who are confident in their academic abilities monitor their work time more effectively, are more efficient problem solvers, and show more persistence than do equally able peers with low self-efficacy" (Usher & Pajares, 2008, p. 751). However, self-efficacy and its impact on performance are heavily influenced by context, and important questions remain unanswered in the literature about the nature and sources of self-efficacy among students in advanced mathematics courses.

The goal of this study was to add to the body of social cognitive research in mathematics education by helping to clarify the roles self-efficacy and calibration play in the mathematical performance of students in a secondary mathematics teacher preparation program. This spurred a thorough review of related literature, development of a model for advanced mathematics performance based on the social cognitive learning theory, and a mixed methods research design that blended broad statistical trends with qualitative themes from task-based interviews. Based on cross-sectional survey and exam performance data for 195 students enrolled in 12 classes ranging from Calculus I to Probability, analysis of variance and structural equation modeling converged with thematic coding of interviews with 10 prospective secondary mathematics teachers to suggest answers to seven research questions, given below:

Q1 Does high school mathematics achievement have a significant effect on the amount of mathematics in participants' college major?

Q2 Do high school mathematics achievement and the amount of mathematics in participants' college major have significant effects on participants' calibration?

Q3 Do high school mathematics achievement, the amount of mathematics in participants' college major, and calibration have significant effects on participants' self-efficacy?

Q4 Do high school mathematics achievement, the amount of mathematics in participants' college major, calibration, and self-efficacy have significant effects on participants' performance on exams in advanced mathematics?

Q5 Are there significant differences in self-efficacy, calibration, the amount of mathematics in participants' college major, and advanced mathematics performance by participants' gender?

Q6 Are there significant differences in self-efficacy and calibration by item difficulty?

Q7 In what ways do prospective secondary mathematics teachers' mathematical problem-solving compare and contrast with the hypothesized relationships between self-efficacy, calibration, and performance in advanced mathematics?

This chapter contains a summary of the study, with special emphasis on situating the research in the context of related literature, consideration of the scope and limitations of the findings, and recommendations for future research. The narrative initially focuses on the research design and methodology before moving to a summary of the quantitative and qualitative findings, which is then followed by a discussion of some implications of the study, limitations of the research, and suggestions for follow-up investigations to extend and clarify the results.

Summary of the Study

The cross-sectional study described in this dissertation employed a mixed methods research design in which task-based interviews with 10 secondary mathematics majors were used to contextualize and triangulate findings gleaned from the quantitative data sources – background surveys, registration data, self-efficacy surveys in the minutes just before final exams, and photocopies of final exams for 195 participants. The setting was the mathematics department at a single mid-sized U.S. doctoral granting university in the Mountain West which specialized in the preparation of secondary mathematics teachers. Data were collected in the last eight weeks of the Spring 2009 semester, and participants were enrolled in at least one of 12 selected mathematics classes offered as part of the requirements to obtain a secondary mathematics major at the research site. Classes included Calculus I, Calculus II, Linear Algebra, Discrete Mathematics, Calculus III, Abstract Algebra II, and Probability.

The conceptual framework supporting the research design and data analysis was built primarily from the constructs of social cognitive theory, especially self-efficacy and calibration, with connections to a cognitive science perspective on calibration as well as path models of mathematics performance developed by Pajares and Kranzler (1995) and Chen (2002). Two pilot studies (detailed in Chapter I) and the review of literature suggested rationale for a structural equation modeling approach to investigating mathematics self-efficacy and calibration among college students, along with a need for a mixed methods inquiry of self-efficacy and exam performance to consider the varied sources of self-efficacy for students who complete advanced undergraduate mathematics. Much of the literature review and research questions rested on the development of a hypothesized structural path model for undergraduate mathematics performance, given as Figure 17, which posited potential direct and indirect effects among high school mathematics achievement, the amount of mathematics in students' college major, mathematics self-efficacy. calibration bias, and final exam performance.

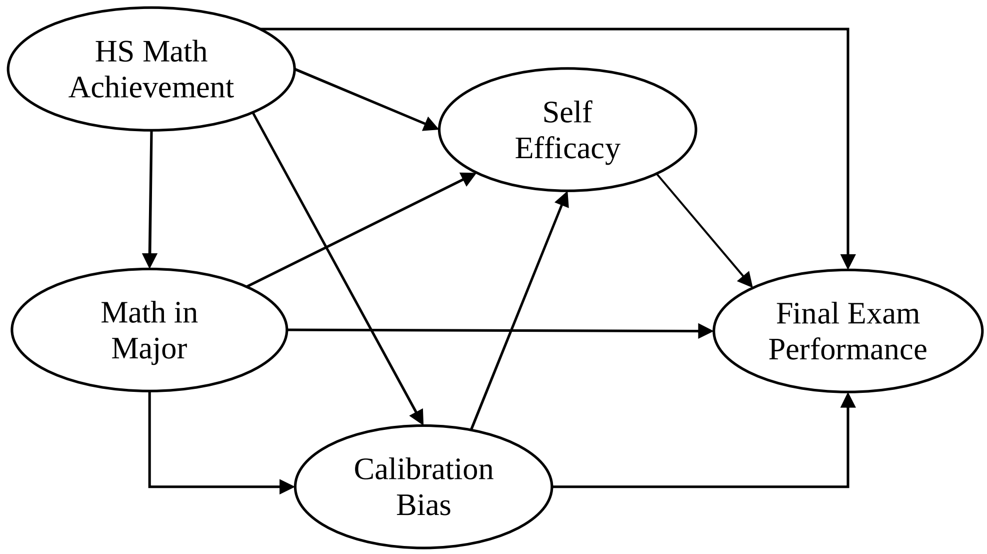

Figure 17. Hypothesized structural path model for advanced mathematics performance.

The data analysis methods employed analysis of variance techniques to test for differences in the measures of the endogenous constructs, self-efficacy, math in major, calibration bias, and final exam performance by gender (Q5) and to test for differences in self-efficacy, calibration bias, and final exam performance by item difficulty (Q6). The first four quantitative research questions (Q1-Q4) were addressed through structural equation modeling, which included the decomposition of observed correlations among indicators of high school mathematics achievement (i.e., self-assessment of high school mathematics performance, high school grade point average, ACT mathematics score) and indicators of self-efficacy, calibration bias, and final exam performance from pre-exam surveys and copies of students' work on final exams. Finally, transcripts and artifacts from the 10 task-based interviews were coded using the conceptual framework, with a special emphasis on the context surrounding and variety of secondary mathematics majors' performance, self-efficacy, and calibration in mathematics. This led to five qualitative themes, which were supported by interview excerpts, vignettes, and examples of student reasoning.

Summary of Findings

In the preceding chapter, the narrative included detailed descriptions of the statistical and qualitative evidence supporting answers to the seven research questions. The aims of this section are to synthesize those findings and to serve as a reference point for further discussion of the meaning, scope, and limitations of the findings.

The primary quantitative findings arose from a structural equation modeling approach to the first four research questions. Starting from the hypothesized structural path diagram (Figure 17), the structural equation model initially included five latent constructs: high school mathematics achievement, math in major, calibration bias, self-efficacy, and final exam performance. However, the number of required mathematics credits associated with students' college major, which served as the single indicator of math in major, was removed from the model because of evidence the data failed several correlation and distribution assumptions of structural equation modeling. Consequently, the final estimated structural model incorporated four latent constructs.

Participants' ACT Mathematics scores, high school grade point average, and self-assessment of their high school mathematics performance served as three indicators of the single exogenous construct, high school mathematics achievement. The two latent constructs from social cognitive theory, self-efficacy and calibration, were each indicated by seven measures based on self-efficacy ratings on surveys just prior to final exams and performance on matched final exam items. Students' performance on seven other final exam items were used as indicators of final exam performance and were randomly selected from quantile-groups of within-class means to represent students' performance on final exam items from a range of difficulties. Collectively, the final estimated structural equation model included 24 indicators spread across four latent constructs.

The estimated structural equation model suggested that each of the indicators loaded significantly onto its specified construct at the .05 criterion. Standardized direct effects among the latent constructs, shown in Figure 18 along with the estimated percentage of variation in the latent constructs explained by the model, suggested the largest direct effect was that of calibration bias on final exam performance ($\beta = -.75$). Calibration bias had a small positive effect on self-efficacy ($\beta = .39$), suggesting a tendency toward overconfidence was more prevalent among students with high self-efficacy than among those with lower self-efficacy. As expected by the review of literature, self-efficacy had a positive effect on final exam performance ($\beta = .62$). Though high school mathematics achievement had a positive effect on self-efficacy ($\beta = .54$), it had an almost equal negative effect on calibration bias ($\beta = -.46$), suggesting increased high school mathematics performance was associated with both increased self-efficacy in college mathematics and a reduction in the tendency toward overconfidence.

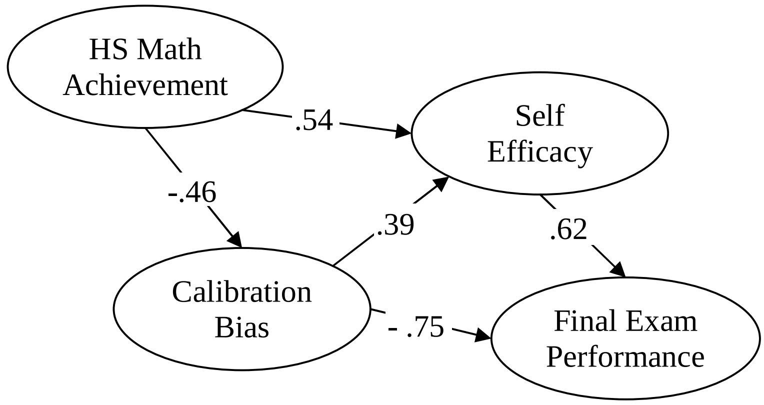

Figure 18. Standardized direct effects in the final estimated structural path model. Indirect effects can be found by multiplying coefficients along multiple-edge paths.

The model did not identify a significant direct effect of high school mathematics achievement on final exam performance, but the effects of high school mathematics achievement on self-efficacy and calibration bias suggested indirect influences on final exam performance. Similarly, the small positive effect of calibration bias on self-efficacy resulted in an indirect effect of calibration bias on final exam performance (β = .24), thus mitigating the large negative direct effect of calibration bias on final exam performance so that the total effect of calibration bias on final exam performance was moderately negative (β = -.51) . After combining direct and indirect effects, high school mathematics achievement had a weak positive effect on self-efficacy (β = .36), a moderate negative effect on calibration bias (β = -.46), and a moderate indirect positive effect on final exam performance (β = .57).

The fifth research question (Q5), was addressed by multiple analysis of variance (MANOVA) tests for potential differences by gender in the composite measures of self-efficacy, calibration bias, math in major, and final exam performance. While the statistical evidence supported the claim that the data met the assumptions of MANOVA, the omnibus test for differences by gender in the composite scales of math in major, self-efficacy, calibration, and final exam performance by gender was not significant (Wilk's $\Lambda = .97$, $F(4, 172) = 1.3$, $p = .27$). That is, there was insufficient evidence to support differences in any of the composite measures by gender.

A post-hoc analysis of the required mathematics in students' majors pointed to possible differences by gender in the percentage of students in advanced mathematics who were mathematics majors. This was evidenced by the observation that 79% (34/43) of female mathematics majors chose the secondary teaching emphasis compared to just 37% (11/30) of the male mathematics majors. These proportions were significantly different ($\chi^2(1, N = 195) = 13.4$, $p < .001$) and may have contributed to an overall difference in the proportion of mathematics majors by gender ($\chi^2(1, N = 195) = 3.9$, $p < .05$) in which 44% (43/97) of female participants were mathematics majors compared to 31% (30/98) of male participants.

The final quantitative research question (Q6) addressed the extent to which study participants' self-efficacy ratings and calibration scores differed according to the difficulty of the exam items represented on the pre-final exam surveys. To allow for comparison across sections, the seven tasks presented on each survey were reverse rank-ordered by the percentage of students who successfully solved each task. For example, a "Level 1" difficulty rating indicated the "easiest" survey task in the sense that it was correctly solved by the highest percentage of students. Applying one-way repeated measures analysis of variance (ANOVA), there were significant main effects of item difficulty on both self-efficacy ratings ($F(6,1164) = 36.6$, $p < .001$) and calibration bias scores ($F(6, 1164) = 14.9$, $p < .001$). That is, both self-efficacy ratings and calibration bias scores tended to decrease with item difficulty.

Post-hoc comparisons of means by item difficulty using Tukey's honestly significant differences (HSD) criterion indicated consistently lower self-efficacy ratings on increasingly difficult items, with self-efficacy means for items separated by at least two levels of difficulty differing in 13 of the 15 possible cases. The mean self-efficacy rating on Level 1 items, for example, was 4.3 (out of 5), while the mean self-efficacy rating on Level 4 items was 3.6 and the mean self-efficacy on Level 7 was 3.0. A similar tendency toward decreased mean calibration bias with increased item difficulty was less consistent than the trend in self-efficacy means. Though the mean calibration bias on the least-difficult items (Levels 1 & 2) were significantly greater than calibration bias on the most-difficult items (Levels 6 and 7), calibration bias means of moderately difficult items (Levels 3-5) were not statistically different.

The final research question (Q7) called for a qualitative inquiry into processes and experiences surrounding the hypothesized relationships among calibration bias, self-efficacy, and performance in college-level mathematics. Analysis of data from task-based interviews with secondary mathematics majors in Calculus I (3 participants), Calculus II (3 participants), and Probability (4 participants) led to five qualitative themes. These included (1) strong perceived high school mathematics performance, (2) lowered self-efficacy following perceived low exam performance, (3) content-specific evaluations of self-efficacy for interview tasks, (4) tendency toward slight overconfidence with improved calibration on low self-efficacy items, and (5) increased self-efficacy to complete a mathematics course after initially not passing the course.

As evidenced by the first qualitative theme, the interview participants generally reported positive perceptions of their high school mathematics performance and preparation for college-level mathematics. Of the 10 interview participants, eight completed classes in high school in which they received college mathematics credit. The two remaining interview participants, Heather and Matthew, each took four years of high school mathematics and began college-level mathematics in Calculus I. With the exception of Nicole, who failed pre-calculus in high school and attributed the poor performance to a lack of effort, the interview participants described high performance in high school mathematics and high mathematics self-efficacy upon high school graduation.

The second interview theme emerged from thematic coding of the interview data using Bandura's (1997) four sources of self-efficacy. Each of the four sources of mathematics self-efficacy were supported by the interviews, but perceived exam performance, especially personal exam scores (mastery experiences) and the perceived exam scores of peers (vicarious experiences), appeared to take a primary role in the development of mathematics self-efficacy. While low perceived exam performance typically led to lowered self-efficacy, participants' feelings about their instructors, especially their approachability, appeared to mediate perceptions of exam performance. In particular, students who reported disliking their instructor on a personal level described self-handicapping behavior that led to low performance, while students who liked their instructor described increased persistence and a willingness to accept low exam performance.

The evidence in support of the second theme was bolstered by vignettes of four Probability students' experiences in Calculus III one year prior to the study. Each of the students, Jennifer, Elizabeth, Emily, and Sarah, independently described similarly low perceived exam scores, but reported a range of overall course grades (B, C, B, and F, respectively). The students' varying reactions to the disappointing exam performance, together with apparent differences in how the exams influenced their self-efficacy, suggested wide-ranging potential for low exam scores to affect secondary mathematics majors' self-efficacy to complete advanced mathematics. Sarah and Emily, for example, both reported doubting their choice to major in mathematics after earning several exam scores below 70% in Calculus III.

The third and fourth qualitative themes, which described students' self-efficacy and calibration on mathematics tasks in an interview setting, served primarily to triangulate and contextualize the larger-scale quantitative findings. When asked to describe their reasoning for selecting self-efficacy ratings on the scale implemented in the study, the participants gave content-specific reasoning, especially recollections of prior attempts at similar problems, familiarity with content, and the perceived number of steps required to solve the problems. Interestingly, none of the participants rated any of the interview tasks with the lowest available self-efficacy rating (1 on a scale of 1 to 6), effectively limiting self-efficacy ratings to a 4 point scale and eliminating the possibility of obtaining a 0 calibration bias score on incorrectly solved items. In addition, five participants expressed aversion to the highest available self-efficacy rating. Taken together, the tendencies to avoid the two extremes on the five-point self-efficacy scale suggested a limitation in the validity of self-efficacy ratings associated with the possibility of response styles.

Qualitative analysis of the interview participants' calibration on interview tasks suggested participants ranged from moderately underconfident (Emily in Probability) to essentially calibrated (Elizabeth and Jennifer in Probability, Nicole and Justin in Calculus II), slightly overconfident (Sarah in Probability, Jackie in Calculus II, Megan in Calculus I), and moderately overconfident (Heather and Matthew in Calculus I). The observed tendency in the interviews toward increased calibration bias of students in the lower-level mathematics course (Calculus I) suggested several plausible explanations, including (1) instructional differences, (2) improved calibration as a result of additional mathematics experience, (3) differences in the relative difficulty of items on the self-efficacy surveys, and (4) chance.

The final qualitative theme considered participants' perceptions of how failing to pass college mathematics classes affected their mathematics self-efficacy. Of the eight participants who had completed at least one college mathematics class prior to the study, five participants had earned an F in at least one college mathematics class. Nicole had failed Calculus I twice, Megan failed Calculus I, Sarah failed Calculus III, Elizabeth failed Discrete Math, and Justin failed Calculus I, Calculus II, Discrete Math, and Probability. Nonetheless, each participant had persisted toward his or her goal to earn a secondary mathematics major, and was enrolled in a college mathematics class at the time of the study.

The analysis of the interview participants' perceptions of failing a college mathematics class suggested, though often initially disappointed in their poor performance, the participants perceived increases in their mathematics self-efficacy after not passing the classes. All five participants said attempts at mathematics classes were easier after the first attempt because of familiarity with course content and a preference for the new instructors. Justin, who had failed four college mathematics classes, reported high overall mathematics self-efficacy and attributed his pattern of failing to pass mathematics classes on the first attempt to a need to “preview” classes. In addition, Nicole believed her history of struggling in Calculus I was going to be an asset as a middle school mathematics teacher. Though the sampling procedure necessarily excluded students who did not choose to persist in their college mathematics coursework after failing to pass one or more classes, for the participants who did persist, the evidence suggested they considered themselves more prepared and more likely to succeed in subsequent attempts at the courses after their initial (non-passing) experience.

Synthesis of Quantitative and Qualitative Findings

The quantitative and qualitative strands of the inquiry, while incorporating differing data sources and analysis techniques, both employed a conceptual framework that focused on high school mathematics performance, self-efficacy, calibration, and exam performance in advanced mathematics. This important aspect of the mixed methods research design afforded opportunities for qualitative themes to triangulate and contextualize the broad scale quantitative findings. The upcoming narrative compares and contrasts the quantitative and qualitative results through the constructs in the structural path model, including comparisons of indications from each strand of the inquiry regarding high school mathematics achievement, self-efficacy, calibration bias, and exam performance.

High School Mathematics Achievement

Both the qualitative and quantitative strands of the investigation suggested students in advanced mathematics classes performed well in high school mathematics. As described in the quantitative results, the participants typically had moderate-to-high scores on each of the three indicators of high school mathematics achievement, including ACT Math scores, high school grade point average, and self-assessment of high school performance on a 7-point ordinal scale. Study participants' average ACT Math score (M = 24.9, SD = 3.9) was about one standard deviation above that of the population of incoming students at the university, and 17% of participants' recorded high school grade point averages were 4.0, compared to just 7% of incoming students at the university. Interestingly, only 1 of the participants rated their high school performance as "Excellent", while 87% (168/ 195) chose one of the descriptors "Very Good", "Good", or "Okay." The qualitative interviews revealed that 8 of the 10 secondary mathematics majors had completed a college mathematics class while in high school, and all described entering college feeling prepared for (at least) Calculus I. Participants reported high self-efficacy in mathematics upon high school graduation, and several participants (e.g., Jennifer, Elizabeth, Jackie) said their college level mathematics was made easier because of their strong high school mathematics preparation.

Self-Efficacy

The qualitative interview data supported the validity of the self-efficacy survey protocol, with interview participants typically describing task- and content-specific reasoning for choosing self-efficacy ratings between 1 and 6, especially familiarity with content, prior experiences with similar tasks, and perceptions of the number of steps required to complete the tasks. Analysis of responses styles suggested that several participants had an aversion to using the highest available rating, and none of the participants chose the lowest available self-efficacy rating. This introduced possible limitations in the effective range of self-efficacy survey data due to the chance that some individuals may have avoided the extremes of the self-efficacy scale. The quantitative analysis of differences in self-efficacy ratings by item difficulty, however, suggested a pronounced pattern of reduced self-efficacy means associated with increased item difficulty, further supporting the validity of the self-efficacy ratings. The estimated structural measurement model also supported the qualitative evidence that self-efficacy ratings reflect task-specific cognitive judgments (as opposed to generalized feelings of confidence), with 69% of the total variation in the seven indicators of self-efficacy left unexplained by a one-factor model.

The structural equation modeling results pointed to a primary role of self-efficacy both as a direct influence on exam performance and as an intermediate influence on effects of high school mathematics achievement and calibration on exam performance. Though calibration bias had the largest direct effect on final exam performance ($\beta = -.75$), self-efficacy had the largest total effect ($\beta = .62$) on final exam performance, exceeding the total effects of both calibration bias ($\beta = -.51$) and high school mathematics achievement ($\beta = .57$). The weak estimated positive effect of calibration bias on self-efficacy ($\beta = .39$) substantiated indications in the pilot studies and review of literature (see Chapter I) that self-efficacy and calibration bias exhibit essentially independent effects on exam performance in mathematics.

The structural equation model incorporated only high school mathematics achievement and calibration bias as sources of variation in mathematics self-efficacy. However, the qualitative interview data helped to contextualize the sources of mathematics self-efficacy through mastery experiences, social persuasions, vicarious experiences, and physical reactions. The interview participants' descriptions of factors which influenced their mathematics self-efficacy focused primarily on mastery and vicarious experiences, with perceived exam scores having the greatest apparent impact on mathematics self-efficacy. Summaries of four students' experiences coping with low perceived exam performance in Calculus III helped to outline the processes supporting exam scores as a source for mathematics self-efficacy, including mediating factors such as social comparisons, the perceived approachability of instructors, and personal like or dislike of instructors.

While the quantitative investigation of mathematics self-efficacy focused on individual tasks representative of exams in the participants' courses, the qualitative inquiry included discussion of more general self-efficacy to pass college mathematics classes (with a C or better). Since 5 of the 8 students who completed at least one college mathematics class prior to the study had earned an F in at least one such class, one qualitative theme described the participants' perceptions of how failing to pass a college mathematics class affected their mathematics self-efficacy. The interview participants reported higher mathematics self-efficacy after failing a college mathematics class. The sources for this increased self-efficacy gleaned from the qualitative analysis included (1) increased familiarity with course content, (2) a perceived improvement in the chances for success with a new instructor, and (3) increased awareness of the personal choices needed to succeed in mathematics.

Calibration Bias

The primary quantitative findings regarding calibration bias included (1) a general tendency toward overconfidence with better calibration associated with more difficult tasks, (2) no significant differences in calibration bias by gender, (3) a large direct negative effect of calibration bias on final exam performance which was mitigated somewhat by a small indirect positive effect on final exam performance through self-efficacy, and (4) high variability in calibration bias scores across exam items. The task-based interview data largely supported the first and fourth of these quantitative findings, while contextualizing the processes that support calibration bias through illustrative examples and providing tentative indications that the calibration bias of secondary mathematics majors may differ by courses.

Calibration bias was operationally defined as the difference between a participant's self-efficacy rating and performance score on a mathematical task, so that positive calibration bias scores were meant to indicate overconfidence. However, the interview data suggested positive calibration bias, especially small positive scores, may be associated with essentially calibrated students. For example, none of the interview participants selected the lowest available self-efficacy rating (1 out of 6), so that every incorrectly solved task in the interviews corresponded to a positive calibration bias score. When asked to attempt two tasks from among those with low or high self-efficacy ratings, only one interview participant correctly completed the low self-efficacy task and half (5 of 10) correctly completed the high self-efficacy task. This meant that 14 of the 20 completed tasks resulted in a positive calibration score and that 9 of the 10 participants obtained a positive combined calibration score on the two items. However, the qualitative analysis of interview participants' reasoning for selecting self-efficacy ratings and subsequent performance suggested four of the students were calibrated. That is, the observed statistical tendency of study participants to obtain positive calibration scores on all but the most difficult tasks, may have, in part, been related to an aversion to choosing the lowest available self-efficacy rating. Follow-up studies could mitigate this threat to the validity of calibration scores by broadening the range of self-efficacy values.

Another connection between the qualitative and quantitative findings about calibration bias arose from the observation that the three interview participants enrolled in Calculus I seemed to be considerably more overconfident than the participants in Calculus II and Probability. Though this could be due to chance, the qualitative finding added context to the observation in the quantitative strand that calibration bias scores showed high variability across items, and suggested the possibility that calibration bias may be influenced by developmental or course-specific processes. In particular, two plausible sources for variation in calibration bias – instructional differences in the respective mathematics classes of the interview participants and development of calibration bias with increased exposure to advanced mathematics classes – were not addressed by the research design.

Final Exam Performance

The research methodology included analysis of final exam performance for 195 students in 12 classes ranging from Calculus I to Probability. In each of the eight separate final exams, seven final exam items were selected for inclusion on self-efficacy surveys and seven items were randomly selected by item difficulty as indicators for final exam performance. This means the analysis included students' performance on a total of 392 authentic final exam tasks. This necessarily introduced variation into the performance data, some of which was accounted for by a variety of measures ranging from estimates of inter-rater reliability, to estimates of uniqueness among the performance indicators in the structural measurement model, to quantile-based sampling of items to ensure representativeness. Nonetheless, the structural equation model explained a remarkably high proportion (83%) of the total variation in the latent construct associated with the final exam performance indicators.

The performance scores earned by participants on the exam items selected as the indicators of final exam performance suggested the final exams included items with a wide range of item difficulty. For example, just 25% of students correctly completed the task selected as the “Level 1” indicator of final exam performance, 58% of students correctly completed the “Level 4” task, and 83% of students correctly completed the “Level 7” task. Using the dichotomous scoring scale, students’ correctly completed a mean of 4.1 of 7 items (*SD* = 1.8), or about 59% of the sampled tasks.

The qualitative inquiry focused primarily on processes surrounding mathematics self-efficacy and calibration in a task-based setting. However, the talk-aloud methodology provided some insight into the validity of a dichotomous scoring (correct/incorrect) scheme for assessing exam performance as well as some potential consequences for the ways in which self-efficacy and calibration bias related to exam performance. As discussed earlier, one consequence of the dichotomous scoring technique (together with the tendency to avoid the lowest available rating on the self-efficacy surveys) was that all incorrect attempts during the interviews corresponded numerically to positive calibration bias scores. On the other hand, the analysis of the interview participants’ performance suggested that 13 of the 14 attempts marked incorrect were the result of substantive conceptual errors (as opposed to numerical or algebraic errors), which supported the validity of the dichotomous scoring system to discern incorrect attempts from essentially correct attempts.

Implications

The purpose of this study was to better understand the roles self-efficacy and calibration play in the mathematical experiences and exam performance of students taking the content courses of a secondary mathematics major. Building on the review of literature and two pilot studies, the study was expected to (1) add to existing self-efficacy research by including an important and often overlooked population of participants, (2) partially fulfill a need for mixed methods studies in social cognitive research, (3) add to research on the mathematical content knowledge and self-beliefs of prospective mathematics teachers, and (4) inform the practice of the mathematical content preparation of prospective secondary mathematics teachers. While the findings are limited in scope by the research design, setting, and data, the study makes substantive contributions toward each of the four goals. In the following sections, the study findings are considered in terms of implications for educational research and the content preparation of preservice secondary mathematics teachers.

Implications for Research

This study adds to existing literature on mathematics self-efficacy and calibration bias in the context of college mathematics, including findings on potential differences in self-efficacy and calibration associated with gender and the difficulty of mathematical tasks. In addition, the mixed methods methodology and structural equation modeling approach to estimating the relative influences of high school mathematics achievement, self-efficacy, and calibration on exam performance, offered opportunities for contextualized findings. The findings helped to both substantiate results from related literature and suggest additional processes that impact self-efficacy, calibration, and exam performance among secondary mathematics majors.

The research design and model for mathematics performance used in this study were based on models of mathematical performance among middle school (Chen, 2003) and high school students (e.g., Pajares & Kranzler, 1995) along with analysis of differences in calibration and self-efficacy associated with gender and item difficulty (e.g., Chen & Zimmerman, 2007). A limitation in the prior path analysis studies that incorporated mathematics self-efficacy and calibration was identified through the assumption in path analysis that predictor variables are perfectly measured by a single measure. This study extended the path analysis techniques to structural equation modeling, which allowed for multiple indicators of the latent constructs (e.g., self-efficacy) in the path model and estimates of the variation both unique to individual indicators and common across indicators of each construct. The concomitant increases in the validity of estimates of directional effects in the structural equation model, together with the incorporation of qualitative data sources, represented the methodological contributions of this study to the literature on mathematics self-efficacy and calibration.

The study findings, along with those of the two pilot studies, supported educational research evidence suggesting that self-efficacy and calibration exhibit approximately equal and opposite effects on mathematics performance. In particular, Chen’s (2003) findings that calibration has a weak effect on self-efficacy and that both self-efficacy and calibration have moderate to strong effects on mathematics performance is supported by this study and the two pilot studies. The magnitude and sign of the standardized coefficients in the structural path model, and even the proportion of total variation in self-efficacy, calibration, and final exam performance explained by the model, were similar to Chen's path analysis results. The similarities between estimates of directional effects, taken in the context of differing settings and measures of mathematics performance, suggested robustness for findings that self-efficacy and calibration have mediating influences on the effect of prior achievement on future performance in mathematics.

This study's findings regarding differences in self-efficacy, calibration bias, and exam performance by gender and item difficulty can be contrasted with Chen and Zimmerman's (2007) cross-national study of self-efficacy and calibration among middle school mathematics students. The results of this study support Chen and Zimmerman's findings that there were no differences by gender in students' calibration bias, self-efficacy, or performance. Similarly, this study supports Chen and Zimmerman's findings that "as items became more difficult, students lowered their self-efficacy beliefs." (p. 230), and both Chen and Zimmerman's study and this study identified a main effect of item difficulty on calibration bias. However, Chen and Zimmerman found that middle school students' calibration bias increased on more difficult items, while the study reported here found students' calibration bias decreased on more difficult items. These contrasting results are likely related to the differing procedures, setting, and measures of self-efficacy and mathematics performance, but the opposite nature of the observed effects suggests reason for further study.

Besides adding to the research on mathematics self-efficacy, calibration bias, and potential differences in each associated with gender and item difficulty, the qualitative strand of this study contributes to the literature on sources of mathematics self-efficacy.

In their comprehensive review of research on the sources of academic self-efficacy, Usher and Pajares (2008) point to the promise of qualitative methods to describe the techniques students use to select among and appraise the many sources of information available to them about their mathematical competencies:

> Qualitative inquiry provides a phenomenological lens through which the development of efficacy beliefs can be viewed, and it can capture the personal, social, situational, and temporal conditions under which students cognitively process and appraise their beliefs and experiences. (p. 784)

Through five qualitative themes, supported by quotations and descriptive accounts of students' mathematical experiences and self-efficacy to complete mathematical tasks, the qualitative strand of the inquiry suggested several processes that can have primary effects on the mathematics self-efficacy of prospective secondary mathematics teachers.

The qualitative themes identified in this study substantiate the primary role of students' perceptions of their mastery experiences in the formation of self-efficacy (Usher & Pajares, 2008). Among the interview participants, perceptions of low exam scores, in particular, was tied to reduced mathematics self-efficacy. However, as evidenced by the descriptions of four students who experienced low perceived exam performance in Calculus III, the repercussions of exam scores on personal self-efficacy appeared to be affected by social comparisons to the perceived performance of peers and personal feelings about the instructor. The many interpretations of similar exam scores suggested a powerful role of interpersonal relationships between students and their instructors, and the evidence supported Zeldin's (2000) contention that successful mathematics professionals developed self-efficacy primarily through performance attainments (e,g., grades, exam scores) and vicarious experiences of peers and family members.

In addition to the implications of the study for research into the sources of self-efficacy of students in advanced mathematics courses, the qualitative inquiry suggested that secondary mathematics majors perceived increased mathematics self-efficacy after earning an F in a college mathematics class. From a social cognitive perspective, failing to pass a college mathematics class introduces several sources for lowered mathematics self-efficacy, including instances of poor performance during mastery experiences such as exams and social comparisons to higher performing peers. Thus, it was somewhat surprising to find that the five participants who had failed a college mathematics class framed those experiences as leading to increased confidence in their abilities to pass the classes in subsequent attempts. This finding was limited by the small number of interview participants, the selection bias introduced by a lack of participants who may have disengaged from mathematics after not passing one or more courses, and the retrospective nature of the participants' accounts of their mathematics self-efficacy. Consequently, the themes in which participants described increased familiarity with course content and beliefs that multiple attempts at courses improved their chances of success, though supported by the data, are probably best characterized as exploratory and preliminary.

Implications for Secondary Mathematics Teacher Preparation

One rationale for the study was the need for holistic description of the mathematical self-efficacy and calibration of prospective secondary mathematics teachers. While educational researchers have contributed robust descriptions of the self-beliefs prospective elementary teachers hold about mathematics (e.g., Harding-DeKam, 2005), little research was identified regarding the mathematics self-efficacy of prospective secondary teachers. This study, with its focus on the mathematics self-efficacy, calibration and performance of secondary mathematics majors, offered a holistic and contextualized description of the strength and accuracy of secondary mathematics majors' beliefs in their mathematical competencies. The findings, though preliminary, suggested secondary mathematics majors tended to (1) experience strong performance in their high school mathematics preparation, (2) draw on content-specific information when evaluating their self-efficacy to complete mathematical tasks, (3) express slight calibration bias in the form of overconfidence to complete exam items, (4) rely on their perceived exam performance and social comparisons to the performance of peers as primary sources of mathematics self-efficacy, and (5) report increased mathematics self-efficacy to complete a college mathematics class after initially not passing the class.

The study findings can be used to inform the design and instruction of content courses in secondary mathematics teacher preparation programs. In particular, the findings suggested several areas of strength among the population of students enrolled in advanced mathematics courses, including prior success in mathematics and moderate to high self-efficacy to learn mathematics. Instructors can draw on this perceived record of accomplishment and self-efficacy by communicating to students that, just as they were able to learn earlier mathematics, the students can expect to succeed in learning new mathematics through persistence and the recognition that increasingly complex content requires increasingly adaptive learning techniques. Based on the review of literature, students' calibration may improve with frequent mastery experiences with moderately difficult tasks, and prompt and clear feedback on the outcomes of performance attempts (O'Connor, 1989). Educational interventions could include "calibration quizzes," whereby students would rate their self-efficacy to complete tasks on a regular quiz, attempt the quiz, and subsequently compare the confidence ratings with their performance on the items, and the effectiveness of such a calibration training (Lichtenstein & Fischoff, 1980) approach could be evaluated through future research. Nonetheless, the review of literature and structural equation modeling findings collectively suggest that improved calibration bias could help secondary mathematics majors develop more accurate perceptions of their mathematical competencies, which in turn is linked to higher self-efficacy and exam performance.

Since the interview participants described a strong reliance on perceived exam performance as a source for overall mathematics self-efficacy, instructors of the content courses for secondary mathematics majors may benefit from clearly communicating their intentions and expectations surrounding exam scores. The interview participants seemed to perceive exam scores below 70% to represent failing scores, so if an instructor has differing perceptions of such scores, the students might benefit from the instructor describing the relative meaning of exam scores as an indicator of understanding or performance. Sarah, who perceived failing exam scores in both Discrete Math and Calculus III, for example, described a higher self-efficacy in Discrete Math because the instructor included a letter grade next to the total score on exams. Especially considering the evidence that students' with lowered self-efficacy in advanced mathematics classes sometimes engaged in self-handicapping behavior that ultimately decreased their chances of passing the classes, students might particularly benefit from clear communication about levels of exam performance that the instructor perceives to be passing or failing.

One indication of overall mathematics self-efficacy from the qualitative strand of the investigation seemed particularly cogent in the preparation of prospective secondary mathematics teachers. Of the eight interview participants who completed at least one mathematics class prior to the study, five participants reported failing to pass a total of nine college mathematics classes. This seemingly high incidence of failed classes within the students' secondary mathematics core content, combined with the perceived benefits the interview participants described for their mathematics self-efficacy, suggests a need for future study. In particular, how often do secondary mathematics majors fail to pass college mathematics classes, and what short-term and long-term effects do such experiences have on their mathematics self-efficacy and career trajectory? These questions are outside the scope of this study, but could prove meaningful in the implementation of secondary mathematics teacher preparation programs, including course sequencing, tracking of students' performance, and advising.

Finally, the interview data suggested instructors played a large role in the interview participants' perceptions of their mathematics self-efficacy. When asked to describe specific qualities of instruction that made them feel more or less confident in their mathematical skills, the students tended to focus on interpersonal skills such as approachability and the apparently intuitive quality of whether the student liked an instructor on a personal level. This exploratory finding suggested future investigation of the ways in which instructional practices are associated with the mathematics self-efficacy of prospective secondary mathematics teachers, especially the qualities of instruction most associated with perceived and observed increases in self-efficacy among mathematics teachers.

Limitations of Study Findings

Considerations of the quality of the research design, including measures to mitigate threats to the internal and external validity of the quantitative dat, as well as efforts to ensure the trustworthiness of qualitative findings, are detailed in the methodology chapter. In addition, consideration of the scope and transferability of findings to other settings and populations was discussed in the methodology chapter along with special emphasis on rich description of the study participants, data collection and analysis strategies. In particular, as a cross-sectional study which focused on students' performance on regular classroom exams, the research design lacked procedures to establish causality among any of the variables. Instead, directional effects among latent variables, together with observed differences by gender and item difficulty, could only describe statistical associations among indicator variables in the context of the review of literature. In the following paragraphs, some additional limitations in the study findings are considered to help contextualize the scope and transferability of the results.

Many of the study findings relate to the structural equation modeling of indicators of high school mathematics achievement, self-efficacy, calibration bias, and final exam performance among data gathered from students in advanced mathematics courses. Although the validity of these findings was strengthened by adequate sample size and application of recommended procedures for model specification and handling of missing data (Schrieber, 2008), the fit indices for the final estimated model suggested only marginal model fit. This, combined with some indications of multidimensionality among indicators in the estimated measurement model, introduced a possibility the estimates of standardized effects among the four latent variables in the model may be vulnerable to

Type I error. One source for these limitations might be non-estimated effects of confounding variables not included in the study, such as participants' academic level, differences in the difficulty of exams, and course-level or instructor effects.

The qualitative and quantitative strands of the inquiry produced largely complimentary findings, and the convergence of themes regarding high school mathematics performance, self-efficacy, calibration, and exam performance had the effect of strengthening the trustworthiness of findings from both strands. However, some limitations were identified during the analysis of the interview data that weakened the quality of the emergent qualitative themes. In particular, the task-based interview data did not include data from sources that may have helped to contextualize the participants' perceptions of their mathematical experiences. Participants described instructional practices, grading policies, and performance of their peers on exams, for instance, but no datum was collected regarding their instructors' perceptions of exam performance or grading policies. These additional data could have added a counter-narrative (Milner, 2007) to the students' descriptions of their experiences which would likely have further contextualized findings and suggested additional insights into the processes supporting mathematics self-efficacy. Classroom observations, as well as interview participants' high school and college mathematics transcripts, could also have helped to triangulate and contextualize the qualitative themes.

Moreover, the discussion of the study findings has included reference to several limitations of the results that emerged from the triangulation of qualitative and quantitative findings. These included (1) the omission of data regarding instructional practices, (2) the possibility of differing roles of mathematics self-efficacy and calibration on mathematical performance tasks other than exams, (3) indications that some participants may be averse to reporting self-efficacy ratings at the two ends of the self-efficacy scale, (4) potential variation in the overall difficulty of exams, and (5) the possibility of longitudinal changes in calibration and self-efficacy during college. In addition, the qualitative strand identified a theme that mathematics self-efficacy can be influenced by failing to pass college mathematics classes, while the quantitative strand did not include any data on students' prior performance in college mathematics classes.

Recommendations for Future Research

The design and interpretations of data in this study were based on decades of educational research into self-efficacy, calibration, and performance, much of which took place in arenas outside of mathematics learning. Consequently, a natural consideration for future research would be the adaptation of the study design and modeling approach to other educational settings. For example, the literature review included Zhao and Linderholm's (2008) review of research into metacomprehension accuracy, a topic that closely aligns with calibration bias, and future research into reading comprehension performance might consider incorporating a social cognitive model like the one used in this study. Besides applications of the conceptual framework or methodology to other educational arenas, the research findings and limitations have suggested several avenues for follow-up research in mathematics education.

In the paragraphs that follow, five follow-up studies are outlined with the goals of inspiring future self-efficacy research in mathematics education and adding to the body of research on how self-efficacy, calibration, and performance interact among students enrolled in advanced mathematics courses. The studies include (1) a larger-scale study of self-efficacy, calibration, and advanced mathematics performance with a single performance measure, (2) a longitudinal inquiry into the trajectories of self-efficacy, calibration bias, and performance of freshman secondary mathematics majors, (3) a cross-sectional investigation of associations between instructional practices and self-efficacy, calibration bias, and exam performance, and (4) a mixed methods inquiry into the effects of failing college mathematics courses on self-efficacy among secondary mathematics majors, and (5) a cross-sectional study of mathematics self-efficacy, calibration bias, and performance across various performance formats.

Though strengthening the transferability of findings to a variety of mathematics content courses, one potentially large source of unexplained variation in the study was the differing exams that served as the basis for self-efficacy, calibration, and performance indicators. A larger-scale study that includes multiple research sites might be able to focus on a single mathematics content course offered at many universities that prepare secondary mathematics teachers, such as Abstract Algebra. The multiple research sites would naturally introduce variation due to the many variations in content across universities, but may also allow for the administration of a single standardized mathematics measure and common self-efficacy surveys across participating sections. Measures would need to be taken to ensure the validity of such a common exam, and much descriptive information would need to be gathered on the students and instructors at the many research sites in order to account for relevant contextual variables. However, the structural equation modeling results could provide additional insights into the generalizability of relationships among self-efficacy, calibration bias, and mathematics performance across research settings under a single measure of performance.

Self-efficacy tends to change with experience in a domain (Bandura, 1997), and characterizing the longitudinal trajectories and processes supporting changes in mathematics self-efficacy throughout college mathematics would represent a substantial addition to mathematics self-efficacy research. In what ways do secondary mathematics majors' self-efficacy, calibration, and performance evolve throughout the students' content preparation? A researcher could address this question using qualitative or mixed methods, starting with interviews of secondary mathematics majors when they first declare their major. With the context of self-efficacy surveys and task-based interviews in successive mathematics courses, the researcher could develop case studies to illustrate the variety of participants' mathematical experiences and the perceived effects of these experiences on mathematics self-efficacy. These data could also be collected as part of efforts to evaluate retention and recruitment in a secondary mathematics program, and the findings could help to identify mathematics classes and experiences which serve to support or diminish participants' self-efficacy and future performance.

Self-efficacy, calibration bias, and mathematics performance may well be affected by both individual's self-beliefs and the instructional practices they experience in college mathematics. Toward that end, future research could include a cross-sectional study of associations among instructional practices and students' self-efficacy, calibration and performance on exams. Using classroom observation data, self-reported descriptions of teaching practices from instructors, course documents, and surveys of students about their perceptions of instruction, a researcher could gather data on instructional practices such as assessment formats, exam difficulty, learning activities, and sources of performance feedback. Statistical techniques could then be used to test for associations between instructional practices and students' self-efficacy, calibration, and exam performance. Though such findings would be preliminary, and multiple research cycles might be necessary to explore the nature of any associations between instruction and mathematics self-efficacy, such research could help to identify ways in which both students and instructors can improve their chances of succeeding toward their goals in the classroom.

One unanticipated, and particularly tentative, finding in the qualitative strand of the inquiry involved secondary mathematics majors' perceptions of failing to pass college mathematics classes. A recommended follow-up study could address the phenomenon of earning an F in one or more of the content courses in a secondary mathematics preparation program from a mixed methods point of view. To what extent do secondary mathematics majors who fail one or more mathematics classes persist toward completing their degree? Answers to these questions could have potential implications for advising secondary mathematics majors and could add to the research on sources of mathematics self-efficacy, especially regarding the relative effects of earning failing grades on the one hand, and perceiving increased familiarity with content on the other. As in the suggested study regarding trajectories of mathematics self-efficacy throughout college, a study on the effects of not passing college mathematics classes could provide a wealth of information through the use of case studies.

Finally, future research could address the potential for contrasting relationships among self-efficacy, calibration, and performance in assessment formats other than regular in-class exams. Two of the mathematics sections offered at the research site during the time of the study chose not to offer traditional open-response timed in-class exams, and it is intriguing to consider the possibility that students' self-efficacy and calibration might take-on a different role in differing performance tasks. A cross-sectional study could address the nature of mathematics self-efficacy in project-based or portfolio assessments, for instance, through methods similar to those employed in the reported study. Instead of completing a self-efficacy survey in the minutes just prior to taking an exam, students might rate their confidence that they can attain the highest mark on a learning outcome listed on a project assignment using the project rubric. If the sample of participants in classes which do not use traditional exams is particularly small, the data collection and analysis could focus on developing emerging understandings through task-based interviews, artifacts, and classroom observation.

References

ACT (2007). *ACT high school profile report: The graduating class of 2007*. Retrieved February 22, 2010, from www.act.org/news/data/07/pdf/National2007.pdf

Alderman, M. K. (1999). *Motivation for achievement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ball, D., & McDiarmid, G. (1989). *The subject matter preparation of teachers* (Issue Paper No. 89-4). Retrieved September 25, 2008, from Michigan State University, National Center for Research on Teacher Learning Web site: http://ncrtl.msu.edu/http/ipapers/html/pdf/ip894.pdf

Bandura, A. (1995). Comments on the crusade against the causal efficacy of human thought. *Journal of Behavioral Therapy and Experimental Psychology, 26*(3), 179-190.

Bandura, A. (1997). *Self-efficacy: the exercise of control*. New York: W. H. Freeman.

Bandura, A., Ross, D., & Ross, S. A. (1963). Imitation of film-mediated aggresive models. *Journal of Abnormal and Social Psychology, 66* (1), 3-11.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*, 825-829.

Betz, N. E., & Hackett, G. (1986). Applications of self-efficacy theory to career development. *Journal of Social and Clinical Psychology, 4*, 279-289.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *British Medical Journal, 314*, 572.

Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist, 34*(3), 139-153.

Bouffard-Bouchard, T. (1990). Influence on self-efficacy on performance in a cognitive task. *The Journal of Social Psychology, 130* (3), 353-363.

Branden, N. (1994). *The six pillars of self-esteem*. New York: Bantam.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, J. R., & Beaudry, J. S. (1998). Gender gap linked to differential socialization for high-achieving senior mathematics students. *The Journal of Educational Research, 91* (3), 140-147.

Campbell, N. K., & Hackett, G. (1986). The effects of mathematics task performance on math self-efficacy and task interest. *Journal of Vocational Behavior, 28*, 149-162.

Chen, P. P. (2002). Mathematics self-efficacy calibration of seventh graders. *Dissertation Abstracts International, 63*(3), 858A. (AAT No. 3047203)

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and individual differences, 14*, 79-92.

Chen, P., & Zimmerman, B. (2007). A cross-national comparison study of self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education, 75*(3), 221-244.

College Board. (2008). *ACT-SAT Concordance Tables*. Retrieved January 15,, 2009, from http://www.act.org/aap/concordance/pdf/report.pdf

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods,* *6*(4), 330-351.

Colosi, L. (1997). Reliability and validity: What's the difference? In *The layman's guide to social research methods*. Retrieved November 5, 2006, from http://www.socialresearchmethods.net/

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches.* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Davenport, E. C., Davison, M. L., Kuang, H., Ding, S., Kim, S. K., & Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal, 35*(3), 497-514.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems, 50*, 1-18.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

Dowling, D. M. (1978). The development of a mathematics confidence scale and its application in the study of confidence in women college students. *Dissertation Abstracts, 39*. (UMI No. AAT 7902111)

Elliot, A. J., & Moller, A. C. (2003). Performance-approach goals: Good or bad forms of regulation? *International Journal of Educational Research, 39,* 339-356.

Ewers, C. A., & Wood, N. L. (1993). Sex and ability differences in children’s math self-efficacy and prediction accuracy. *Learning and Individual Differences, 5* (3), 259-

267.

Fischbein, E. (1987). Investigations in overconfidence. In *Intuition in science and mathematics* (pp. 28-42). Dordrecht: D. Reidel Publishing.

Fitchett, S., King, K., & Champion, J. (in press). Outcomes of math placement at the University of Northern Colorado: An analysis of advising and enrollment data. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research, 18* 382-388.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13*(3), 465-486.

Fox, J. (2009). *Package sem*. Retrieved October 22, 2009, from www.r-project.org

Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences.* (6th ed.). New York: Worth Publishers.

Garson, D. (2006). *Path analysis*. Retrieved February 26, 2007, from North Carolina State University Web site: http://www2.chass.ncsu.edu/

Gelman, A., Hill, J., Yajima, M., Su, Y. S., & Pittau, M. G. (2009). *Package mi*. Retrieved October 22, 2009, from www.r-project.org

Glesne, C. (2006). *Becoming qualitative researchers: An introduction* (3rd ed.). Boston: Pearson Education.

Grice, J. W., & Iwasaki, M. (2007). A truly multivariate approach to MANOVA. *Applied Multivariate Research, 12*(3), 199-226.

Guion, L. A. (2002, September). *Triangulation: Establishing the validity of qualitative* *studies*. Retrieved November 4, 2006, from University of Florida Extension Web site: http://edis.ifas.ufl.edu/

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/performance correspondence. *Journal for Research in Mathematics Education, 20*(3), 261-273.

Hackett, G., Betz, N. E., O'Halloran, M. S., & Romac, D. S. (1990). Effects of verbal and mathematics task performance on task and career self-efficacy and interest. *Journal of Counseling Psychology, 37* (2), 169-177.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/?performance measures. *Review of Educational Research, 52* (1), 123-142.

Harding-DeKam, J. L. (2005). Construction and validation of an instrument for assessing prospective elementary teachers' attitudes and beliefs in mathematics. *Dissertation Abstracts International, 66,* 1293A. (UMI No. 3171928)

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42* (2), 371-406.

Husman, J., Brem, S., & Duggan, M. A. (2005). Student goal orientation and formative assessment. *Academic Exchange Quarterly, 9* (3), 355–359.

James, W. (1890). The principles of psychology. In *Classics in the History of Psychology*. Retrieved January 2, 2009, from http://psychclassics.yorku.ca/James/Principles/index.htm

Joreskog, K., & Sorbom, D. (2008). LISREL for Windows (Version 8.8) [Computer software]. Lincolnwood, IL: Scientific Software International.

Kahan, J. A., Cooper, D. A., & Bethea, K. A. (2003). The role of mathematics teachers' content knowledge in their teaching: A framework for research applied to a study of student teachers. *Journal of Mathematics Teacher Education, 6*, 223-252.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Klassen, R. M. (2006). Too much confidence? The self-efficacy of adolescents with learning disabilities. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 181-200). Greenwhich, CT: Information Age Publishing.

Lapan, R. T., Shaughnessy, P., & Boggs, K. (1996). Efficacy expectations and vocational interests as mediators between sex and choice of math/science college majors: A longitudinal study. *Journal of Vocational Behavior*, *49*, 277-291.

Lent, R. W., Brown, S. D., & Hackett, B. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79-122.

Lent, R. W., Lopez, F. G., & Bieschke, K. J. (1991). Mathematics self-efficacy: Sources and relation to science-based career choice. *Journal of Counseling Psychology, 38* (4), 424-430.

Lent, R. W., Lopez, F. G., Brown, S. D., & Gore, P. A. (1996). Latent structure of the sources of mathematics self-efficacy. *Journal of Vocational Behavior, 49*, 292-308.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149-171.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

Lightsey, R. (1999). Albert Bandura and the exercise of self-efficacy [Review of the book *Self-efficacy: The exercise of control*]. *Journal of Cognitive Psychotherapy, 13* (2), 158-166.

Lin, L., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*, 345-391.

Loehlin, J. C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lutzer, D. J. (2002). *Mathematics majors 2002*. Williamsburg, VA: College of William and Mary, Department of Mathematics.

Lutzer, D. J., Rodi, S. B., Kirkman, E. E., & Maxwell, J. W. (2007). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2005 CBMS survey*. American Mathematical Society.

Madewell, J., & Shaughnessy, M. F. (2003). An interview with Frank Pajares. *Educational Psychology Review, 15* (4), 375-397.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of*

Educational Psychology, 97 (4), 723-731.

McDonald, R. P., & Ho, M. R., (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

McMillan, J. H., Singh, J., & Simonetta, L. G. (2001, Winter). The tyranny of self-oriented self-esteem. *Educational Horizons*, 92-95.

Mels, G. (2006). *LISREL for Windows: Getting started guide*. Lincolnwood, IL: Scientific Software International.

Merriam, S. B. (1998). *Qualitative research and case study applications in education* (2nd ed.). San Francisco: Jossey-Bass.

Mertens, D. M. (2005). *Research and evaluation in educational psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

Midgley, C., & Urdan, T. (2001). Academic self-handicapping and achievement goals: A further examination. *Contemporary Educational Psychology*, 26, 61-75.

Milner, H. R. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher, 36*(7), 388-400.

Moncur, M. (2007). Henry Ford (1863-1947). In *Quotations by author*. Retrieved January 2, 2009, from http://www.quotationspage.com/?quotes/?Henry_Ford/

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13* (2), 125-145.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being better and worse than we think we are. *Journal of Personality and Social*

Psychology, 92 (6), 972-989.

Morton, B. A., Peltola, P., Hurwitz, M. D., Orlofsky, G. F., Strizek, G. A., & Gruber, K. J. (2008). *Education and certification qualifications of departmentalized public high school-level teachers of core subjects: Evidence from the 2003-04 Schools and Staffing Survey* (NCES No. 2008-338). Washington, D.C.: National Center for Education Statistics.

Mura, R. (1987). Sex-related differences in expectations of success in undergraduate mathematics. *Journal for Research in Mathematics Education, 18*(1), 15-24.

No Child Left Behind Act, 20 U.S.C. § 6301 (2001).

O'Brien, V., Kopala, M., & Martinez-Pons, M. (1999). Mathematics self-efficacy, ethnic identity, gender, and career interests related to mathematics and science. *Journal of Educational Research, 92* (4), 231-235.

O'Connor, M. (1989). Models of human behaviour and confidence in judgment: A review. *International Journal of Forecasting, 5*, 159-169.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple linear regression that researchers should always test. *Practical Assessment, Research,* & *Evaluation, 8*(2). Retrieved January 15, 2009 from http://PAREonline.net

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24,* 124-139.

Pajares, F., & Kranzler, J. (1995, April). *Role of self-efficacy and general math ability in mathematical problem-solving: A path analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

(ERIC Document Reproduction Service No. ED387342)

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86* (2), 195-203.

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *Journal of Experimental Education, 65*(3), 213-229.

Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. In R. Riding & S. Rayner (Eds.), *Perception* (pp. 239-266). London: Ablex Publishing.

Pajares, F., & Schunk, D. H. (2002). Self and self-belief in psychology and education: An historical perspective. In J. Aronson (Ed.), *Improving academic achievement* (pp. 5-25). New York: Academic Press.

Pajares, F., & Urdan, T. (Eds.) (2006). *Self-efficacy beliefs of adolescents*. Greenwhich, CT: Information Age Publishing.

Patton, M. Q. (2002). *Qualitative research* & *evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Philippou, G. N., & Christou, C. (1998). The effects of a preparatory mathematics program in changing prospective teachers' attitudes towards mathematics *Educational Studies in Mathematics, 35*(2), 189-206.

Revelle, W. (2009). *An introduction to psychometric theory with applications in R.* Retrieved, December 19, 2009, from http://www.personality-project.org/r/book/

Reyes, L. H. (1984). Affective variables and mathematics education. *The Elementary*

School Journal, 84 (5), 558-581.

Rosenthal, J. S. (1995). Active learning strategies in advanced mathematics classes. *Studies in Higher Education, 20* (2), 223-229.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9*, 321-332.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*, 113-125.

Schraw, G., Polenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Schreiber, J. B. (2008). Core reporting practices in structural equation modeling. *Research in Social and Administrative Pharmacy, 4*, 83-97.

Schunk, D. H. (2004). *Learning theories: An educational perspective* (pp. 285-289, 447-451). Upper Saddle River, NJ: Pearson Education.

Schwandt, T. A. (2001). *Dictionary of qualitative inquiry* (2nd ed.). Thousand Oaks, CA: Sage.

Seidman, I. (1998). *Interviewing as qualitative research* (2nd ed.). New York: Teachers College Press.

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

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268.

Simon, S. D. (1999). *From neo-behaviorism to social constructivism? The paridigmatic non-evolution of Albert Bandura*. Unpublished master's thesis, Emory University. Retrieved March 28, 2007, from http://www.des.emory.edu/

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, IA: Iowa State University Press.

Stage, F. K., Carter, H. C., & Nora, A. (2004). Path analysis: An introduction and analysis of a decade of research. *The Journal of Educational Research, 98*(1), 5-12.

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences*, 3rd Ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12* (4), 437-475.

Suhr, D. (2008). Step your way through path analysis. *Western Users of SAS Software Conference Proceedings*. Retrieved January 13, 2009, from http://www.wuss.org/

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129-160.

Thompson, A. G. (1984). The relationship of teachers' conceptions of mathematics and mathematics teaching to instructional practice. In T. Carpenter, J. Dossey, & J. Koehler, *Classics in mathematics education research* (pp. 173-184). Reston, VA: National Council of Teachers of Mathematics.

Urdan, T. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology, 96*(2), 251-264.

Urdan, T., Pajares, F., & Lapin, A. Z. (1997). *Achievement goals, motivation, and performance: A closer look*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED412268)

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future research. *Review of Educational Research (78)*(4), 751-796.

Zeldin, A. L. (2000). *Sources and effects of the self-efficacy beliefs of men with careers in mathematics, science, and technology*. Unpublished doctoral dissertation, Emory University, Atlanta, GA.

Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal, 37*(1), 215-246.

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review, 20* (2), 191-206.

Zimmerman, B. J., & Schunk, D. H. (1989). *Self-regulated learning and academic achievement: theory, research, and practice.* New York: Springer-Verlag.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29* (3), 663-676.

APPENDIX A

Informed Consent Letter

Informed Consent for Participation in Research

Project Title: The Mathematics Self-Efficacy and Calibration of Students in a Secondary Mathematics Teacher Preparation Program

Lead Researcher: Joe Champion, School of Mathematical Sciences, 970-351-2229
Research Advisor: Robert Powers, Mathematical Sciences, 970-351-1157

I am researching the self-confidence and performance of students in advanced mathematics courses. Much of the data I plan to use will come from photocopies of your regular in-class exams. However, I will ask you to complete a background survey and one or more 3-5 minute surveys throughout the semester. In addition, you may also be invited to participate in a 45-60 minute interview where you'll explain your thinking while attempting problems related to the mathematics in your class.

The main questions I'll ask you are about your perception of whether you can complete certain mathematics problems related to your class. These surveys will be administered in the few minutes just prior to your major exams, including your final exam. Any surveys and interviews you complete will take no more than a total of 90 minutes. If you decide to participate in an interview, your work on math tasks and responses to interview questions will be recorded, and the digital audio recordings will be disposed of within 2 years of the date of the interview.

The risks of participation in the study are likely no greater than those associated with taking a college mathematics course, completing background surveys, and working on math problems in a one-on-one interview setting. However, you may experience some anxiety from completing a short survey just prior to a major exam, and if you are concerned about this anxiety you may decline participating in the study at any point. If you choose to participate, you may improve in your ability to estimate your understanding in math and may experience increased awareness of how your beliefs about your math skills are related to your performance in advanced math courses.

Nonparticipation or withdrawal from the study will not affect your grade in the course. Your teacher will not know who in the class is participating. If you do choose to participate, you will not be identifiable in the final report of the study.

Participation is voluntary. You may decide not to participate in this study and if you begin participation you may still decide to stop and withdraw at any time. Your decision will be respected and will not result in loss of benefits to which you are otherwise entitled. Having read the above and having had an opportunity to ask any questions, please sign below if you would like to participate in this research. A copy of this form will be given to you to retain for future reference. If you have any concerns about your selection or treatment as a research participant, please contact the Sponsored Programs and Academic Research Center, Kepner Hall, University of Northern Colorado Greeley, CO 80639; 970-351-1907.

_______________________ _______________________ __________

Participant's Name (please print) Participant's Signature Date

_______________________ 3/9/2009

Researcher's Signature Date

APPENDIX B

Background Survey

Name (Print): ________________________________ Bear #: ____________________ ☐ (don't remember)

1. What is your primary college MAJOR? (Circle One)

Biology	Mathematics
Business	Physics
Chemistry	Pre-Professional, specify: ________________________________
Computer Science	Elementary Teaching (IDLA), emphasis: __________________
Earth Sciences	Undeclared, leaning towards: ____________________________
	Other, specify: ______________________________________

2. What is your academic level at UNC?

   Freshman Sophomore Junior Senior Graduate

3. Does your major include an emphasis in education?

   Yes No

   If *Yes*, which grade band are you MOST interested in teaching? (Circle One)

   Early Childhood Elementary Middle Secondary K-12

4. What is your gender?

   Male Female

5. Complete ALL of the following. On a scale from 0% (unsure) to 100% (completely sure), how confident are you that you can earn the following overall grades in this class?

   I am ____ % sure I can earn a D or better in this class this semester.

   I am ____ % sure I can earn a C or better in this class this semester.

   I am ____ % sure I can earn a B or better in this class this semester.

   I am ____ % sure I can earn an A in this class this semester.

6. How many semesters of mathematics did you complete in high school? ________ semesters

7. Which of the following best describes how well you did in your high school math courses? (Circle One)

   Excellent Very Good Good OK Not So Good Bad Really Bad

8. Circle the listings that best correspond to the math courses you completed in high school.

General Math/Consumer Math	Integrated Mathematics 1	Calculus
Basic Math 1, 2, 3, or 4	Integrated Mathematics 2	AP Calculus
Pre-Algebra	Integrated Mathematics 3	Differential Equations
Informal Geometry	Trigonometry	College Algebra
Geometry	Trigonometry & Geometry	Linear Algebra
Algebra 1	Trigonometry & Algebra	Statistics
Algebra 2	Analysis	Probability
Algebra 3	Pre-Calculus	Probability & Statistics

APPENDIX C

Self-Efficacy Surveys for Final Exams

PRE FINAL EXAM SURVEY – SECTION 1

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

Rating (1-6):

1. If $f(x) = \sqrt{x+3}$, find the equation of the tangent line at $x = 1$.

Rating (1-6):

2. If $x^2 + xy + y^2 = 19$, find the value of $\frac{dy}{dx}$ at the point (2,3).

Rating (1-6):

3. The position function of a particle moving in a straight line is $s(t) = -16t^2 + 48t + 100$, where $s(t)$ is measured in feet and t is measured in seconds. Find the velocity at $t = 2$.

Rating (1-6):

4. A rectangular storage container with an open top is to have a volume of 10 cubic meters. The length of its base is twice the width. Material for the base costs $10 per square meter. The material for the sides costs $6 per square meter. Find the cost of the materials for the cheapest container.

Rating (1-6):

5. Evaluate the following limit

$$\lim_{x\to 2} \frac{x^2 - 4}{\mathrm{x}^2 + 3\mathrm{x} - 10}$$

Rating (1-6):

6. Differentiate $y = 5e^x \cdot \sin(3x + 2)$.

Rating (1-6):

7. If $f(x) = \sqrt{x+1}$, find the equation of the tangent line at $x = 3$.

Rating (1-6):

8. The altitude of a triangle is increasing at a rate of 1 cm/min while the area of the triangle is increasing at a rate of 2 cm^2/min. At what rate is the base of the triangle changing when the altitude is 10 cm and the area is 100 cm^2?

PRE FINAL EXAM SURVEY- SECTIONS 2 & 3

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

Rating (1-6):

1. Differentiate $h(x) = (x-2)(2x+3)$.

Rating (1-6):

2. If $x^2 + xy + y^2 = 19$, find the value of $\frac{dy}{dx}$ at the point (2,3).

Rating (1-6):

3. Differentiate the following:

$$y = 5e^x \cdot \sin(3x+2)$$

Rating (1-6):

4. If $F(x) = f(g(x))$, where $f(-2) = 8$, $f'(-2) = 4$, $f'(5) = 3$, $g(5) = -2$, and $g'(5) = 6$, find $F'(5)$.

Rating (1-6):

5. Find the limit:

$$\lim_{x \to 0} \frac{x + \tan x}{\sin x}$$

Rating (1-6):

6. On what interval is the function $f(x) = x^3 e^x$ increasing.

Rating (1-6):

7. Differentiate $g(x) = (x-5)(3x+1)$.

Rating (1-6):

8. Find the intervals on which $f(x) = x^3 - 12x + 1$ is increasing or decreasing.

PRE FINAL EXAM SURVEY – SECTIONS 4 & 5

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

Rating (1-6):

1. Find the Taylor polynomial of degree 4 of $\cos x$ at $a = \frac{\pi}{3}$

Rating (1-6):

2. Find $\int x^2(x^3+5)^4 dx$.

3. Determine whether the following sequence converges or diverges

$$a_n = \frac{\sqrt{n}}{1+\sqrt{n}}$$

Rating (1-6):

Rating (1-6):

4. Determine if the series $\sum_{n=1}^{\infty} \frac{(3)^n}{5^{n-1}}$ converges or diverges

5. Find the radius of convergence and the interval of convergence for the series:

Rating (1-6):

$$\sum_{n=1}^{\infty} \frac{(x+2)^n}{n \cdot 4^n}$$

Rating (1-6):

6. Integrate $\int \frac{x-1}{x^2+3x+2} dx$ using partial fractions.

Rating (1-6):

7. Find the Taylor polynomial of degree 4 of $\sin x$ at $a = \frac{\pi}{3}$.

Rating (1-6):

8. Integrate $\int x \cdot \ln x \, dx$.

PRE FINAL EXAM SURVEY- SECTION 6

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 2 3 4 5 6

(not sure) (completely sure)

Name (Print):

Rating (1-6):

1. Let A, B, C and D be invertible *n x n* matrices. Solve $AB(X + C)D^{-1} = I_n$ for X.

Rating (1-6):

2. Compute the determinant of the following matrix by cofactor expansion (without using your calculator). Show all work. $\begin{bmatrix} 8 & 2 & 2 \\ 9 & 5 & 8 \\ 3 & 2 & 7 \end{bmatrix}$

Rating (1-6):

3. Find the standard matrix of the transformation T: $R \rightarrow R$ that reflects points about the x-axis followed by a rotation of $\pi/2$ radians in the clockwise direction. Show your work.

Rating (1-6):

4. Let $v_1 = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}$, $v_2 = \begin{bmatrix} -3 \\ -4 \\ 1 \end{bmatrix}$, $v_3 = \begin{bmatrix} 2 \\ 1 \\ 6 \end{bmatrix}$, and $b = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}$. Is b a linear combination of the vectors v_1, v_2, and v_3? Explain why or why not.

Rating (1-6):

5. Let $A = \begin{bmatrix} 5 & -8 & 1 \\ -7 & 2 & -6 \end{bmatrix}$ and let $\boldsymbol{u} = \begin{bmatrix} 8 \\ 2 \\ -2 \end{bmatrix}$. Define a transformation $T: R^m \rightarrow R^n$ by $T(\boldsymbol{x}) = A\boldsymbol{x}$. Find $T(\boldsymbol{u})$, the image of $\boldsymbol{u}$ under the transformation T.

Rating (1-6):

6. Without using your calculator, find the eigenvalues of the matrix $A = \begin{bmatrix} 3 & 3 \\ -2 & -4 \end{bmatrix}$. Show all work.

Rating (1-6):

7. Let A, B, C and F be invertible *n x n* matrices. Solve $AC(X + B)F^{-1} = I_n$ for X.

Rating (1-6):

8. Let W be the set of all vectors of the form $\begin{bmatrix} a + 3b \\ 4b \\ 5a - b \\ -a \end{bmatrix}$ where a and b are arbitrary real numbers. If W is a vector space, find a set of vectors that spans it. Otherwise explain why W is not a vector space.

PRE FINAL EXAM SURVEY – SECTIONS 7 & 8

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

1. Write the coefficient of x^{16} for the expression
$$(x+2)^{30} + x^6(x+5)^{27}$$
Rating (1-6):

2. Find the exact value of $13 + 20 + 27 + 34 + {}_{\dots} + 7286$. Show all steps.
Rating (1-6):

3. Suppose that $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$ and $B = \{a, b, c, d\}$. How many functions are there from B to A that are NOT one to one?
Rating (1-6):

4. Find the exact value of $16 + 12 + 9 + \frac{27}{4} + {}_{\dots} + \frac{3^{41}}{4^{39}}$. Show all steps.
Rating (1-6):

5. Write the contrapositive of the following statement: If a day has the largest amount of daylight for the year, then that day occurs in June.
Rating (1-6):

6. How many bit strings of length 25 and weight 12 DO NOT start with the sequence 111 or end with the sequence 101?
Rating (1-6):

7. Write the coefficient of x^{21} for the expression
$$(x+3)^{37} + x^8(x+2)^{30}$$
Rating (1-6):

8. State whether the following function is one-to-one and/or onto, and explain: $f: B^4 \to B^4$ by $f(a_1, a_2, a_3, a_4) = (a_2, a_3, a_4, a_2)$. For example, $f(1001) = 0010$.
Rating (1-6):

PRE FINAL EXAM SURVEY- SECTIONS 9 & 10

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

1. Let $f(x,y) = \sin(2x+y)$. Find $f_{xy}\left(\pi, \frac{\pi}{2}\right)$.

   Rating (1-6):

2. Let $\vec{w} = \vec{\imath} - \vec{\jmath} + 2\vec{k}$, and $\vec{u} = 2\vec{\imath} - a\vec{\jmath} + 3\vec{k}$. Find the value of a making $\vec{w}$ and $\vec{u}$ perpendicular.

   Rating (1-6):

3. Let $z = e^x \sin y$ and let x and y be functions of s and t with $x(0,0) = 0$, $y(0,0) = 0$, $\frac{\partial x}{\partial s} = 3$ and $\frac{\partial y}{\partial s} = 4$ at $(s,t) = (0,0)$. Find $\frac{\partial z}{\partial s}$ when $(s,t) = (0,0)$.

   Rating (1-6):

4. Let $f(x,y,z) = x^2y + y^3z + xz^3$ and let $P = (2,1,-1)$. What is the maximum rate of change of f at P?

   Rating (1-6):

5. Sketch the region of integration and evaluate :

$$\int_0^9 \int_{\sqrt{y}}^3 \sin(\pi x^3)\, dxdy$$

   Rating (1-6):

6. Convert the following integral to polar coordinates and evaluate it:

$$\int_0^2 \int_{-\sqrt{4-x^2}}^0 \frac{xy}{\sqrt{x^2+y^2}} dydx$$

   Rating (1-6):

7. Let $g(x,y) = \cos(2x+y)$. Find $g_{xy}\left(\pi, \frac{\pi}{2}\right)$.

   Rating (1-6):

8. For $f(x,y) = 2x^3 + y^2 - 6x + 4y$, find and classify the local extrema of f.

   Rating (1-6):

PRE FINAL EXAM SURVEY – SECTION 11

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1	2	3	4	5	6
(not sure)				(completely sure)	

Name (Print):

Rating (1-6):

1. Show that $-\frac{1}{2}+\frac{\sqrt{3}}{2}i$ is an algebraic number.

Rating (1-6):

2. Let S be the subset $\{\bar{0}, \bar{2}, \bar{4}\}$ of the ring $\mathbb{Z}_6$. Make the addition and multiplication tables for S.

Rating (1-6):

3. Find all the roots in $\mathbb{C}$ of the polynomial $q(x) = 2x^3 + x^2 + x - 1$.

Rating (1-6):

4. Is the polynomial $q(x) = 2x^3 + x^2 + x - 1 \in \mathbb{Q}[x]$ reducible or irreducible. Justify your answer.

Rating (1-6):

5. Give definitions of an integral domain and of an ordered integral domain.

Rating (1-6):

6. Let T consist of all real numbers of the form $a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$ with a, b, c, and d rational. Show that T is a subfield in the field of real numbers.

Rating (1-6):

7. Show that $-\frac{1}{2}-\frac{\sqrt{3}}{2}i$ is an algebraic number.

Rating (1-6):

8. Is $\mathbb{Z}_6$ an ordered integral domain? Justify your answer.

PRE FINAL EXAM SURVEY – SECTION 12

DO NOT ATTEMPT the following problems. Instead, rate how confident you are that *you can correctly* complete the items by choosing any number between 1 and 6:

1 **2** **3** **4** **5** **6**

(not sure) (completely sure)

Name (Print):

1. If you roll two 10-sided dice, what is the probability the sum is 6? Rating (1-6):

2. Let $X \sim b(50, .8)$. Estimate $P(X = 40)$ using a Poisson approximation. Rating (1-6):

3. SAT scores are approximately normally distributed with mean 500 and variance 100^2. If X is the SAT score of a randomly chosen student, find $P(525 < X < 600)$. Rating (1-6):

4. Let $X_1, \ldots, X_n$ be a random sample from a $\Gamma(10, \theta)$ distribution (so α is known to be 10, but θ is unknown). Find the maximum likelihood estimator for θ. Rating (1-6):

5. Assume $X_1, \ldots, X_{25}$ is a random sample from a standard normal distribution and $W = X_1^2 + \cdots + X_{25}^2$. What is the distribution of W? Rating (1-6):

6. A recent poll asked 450 American adults, chosen by random dialing, if they would be willing to pay up to 10% more for electricity if that electricity would be generated by wind instead of coal. Of those surveyed, 285 said yes. Give a 95% confidence interval for the proportion of all American adults that would be willing to pay more for wind-generated electricity. Rating (1-6):

7. If you roll two 10-sided dice, what is the probability the sum is 5? Rating (1-6):

8. Let $X_1, \ldots, X_6$ be a random sample from an $Exp(2)$ distribution, and let $Y = \sum_{i=1}^{6} X_i$. Find the moment generating function for Y. Rating (1-6):

APPENDIX D

Interview Protocol

Task-based Interview Protocol

Thank you for agreeing to participate in this interview. The goal of this interview is to talk to you about what you think about when you're asked about what you can do in a mathematics class. The interview is expected to take between 45 and 60 minutes.

Is it OK if I record our discussion? *[If yes, turn on microphone and repeat the question so it is recorded]* When I write about this interview I'd like to give you a fake name. What pseudonym would you prefer that I use?

[*Take out a 7-question self-efficacy survey based on the current course material.]* This is a quick survey asking you to estimate how confident you are that you can solve certain problems related to your class. There are no right and wrong answers, so just write numbers you feel match how confident you are you can complete the problems correctly.

I see that on statement *[select one of the items with the highest rating]* you wrote ________ on the survey, can you tell me why you decided on that number? *[Repeat for a lowest rated task and a middle-rated task.]*

Can you give me an example of a challenging problem in your class that you would say you are **completely sure** you can solve correctly? [Follow-up: *Why did you choose this problem?]* *[Repeat for a problem in which the student would have low self-efficacy]*

Now I'd like you to try completing some sample problems from your class. It's ok if you can't do the problems right now, so please just try your best.

1. *[Choose a task the student marked with high self-efficacy.]*
2. Can you work through the following problem and tell me what you're thinking as you work?
3. *[As the student works, ask them about any similar problems they've done in class or in previous semesters. E.g., Do you recall doing a problem like this on your test?]*
4. Do you think you solved the problem correctly? Why or why not?

[Repeat steps 1-4 for items marked with medium and low self-efficacy.]

Thank you for working through those problems with me. Now I'd like to talk a little more generally about your class this semester. How is the class going for you?

Can you think of anything about your class this semester that has helped you feel more confident about what you can do in the class?

Similar question. Can you tell me about anything in your class this semester that might have made you feel less confident about what you can do in the class?

What about any other college math classes you've had? Which of the classes do you think left you thinking you were better able to learn a new math topic? In which were you less confident?

That's all I have for now. Do you have any questions for me? Is it all right if I follow up with you if I have any questions about what we talked about today? Thank you for taking the time to talk with me, and good luck in your class.

APPENDIX E

Correlations among Indicators in the Structural Equation Model

Correlations among Indicators in the Structural Equation Model

		1	2	3	4	5	6	7	8
1.	HS Self								
2.	HS GPA	0.15							
3.	ACT Math	0.31	0.52						
4.	SE Level 1	0.15	0.21	0.19					
5.	SE Level 2	0.15	0.03	0.13	0.38				
6.	SE Level 3	0.22	0.09	0.24	0.35	0.33			
7.	SE Level 4	0.21	0.10	0.17	0.37	0.30	0.35		
8.	SE Level 5	0.16	0.02	0.08	0.28	0.26	0.36	0.33	
9.	SE Level 6	0.23	0.13	0.11	0.16	0.14	0.39	0.37	0.38
10.	SE Level 7	0.19	0.23	0.13	0.29	0.16	0.38	0.42	0.34
11.	Bias Level 1	-0.12	0.06	-0.11	0.05	0.01	0.10	0.08	0.15
12.	Bias Level 2	-0.18	-0.27	-0.18	-0.03	0.06	0.12	0.05	0.13
13.	Bias Level 3	-0.23	-0.01	-0.09	0.10	0.10	0.07	0.12	0.01
14.	Bias Level 4	-0.17	-0.16	-0.16	0.02	-0.05	-0.05	-0.03	-0.06
15.	Bias Level 5	-0.24	-0.19	-0.16	0.08	0.02	0.01	-0.05	0.01
16.	Bias Level 6	-0.08	-0.03	-0.12	0.04	-0.07	0.06	0.17	0.10
17.	Bias Level 7	-0.12	-0.16	-0.09	-0.01	0.05	0.07	-0.08	0.08
18.	Perf. Level 1	0.32	0.12	0.08	-0.09	0.03	0.09	0.05	0.17
19.	Perf. Level 2	0.28	0.05	0.26	0.12	0.17	0.26	0.21	0.11
20.	Perf. Level 3	0.20	0.25	0.21	0.18	0.10	0.13	0.15	0.09
21.	Perf. Level 4	0.20	0.26	0.27	0.14	0.11	0.27	0.22	0.20
22.	Perf. Level 5	0.23	0.11	0.25	0.18	0.29	0.31	0.14	0.15
23.	Perf. Level 6	0.18	0.23	0.29	0.15	0.25	0.25	0.19	0.19
24.	Perf. Level 7	0.26	0.25	0.34	0.27	0.18	0.12	0.26	0.14

Correlations among Indicators in the Structural Equation Model (continued)

		9	10	11	12	13	14	15	16
1.	HS Self								
2.	HS GPA								
3.	ACT Math								
4.	SE Level 1								
5.	SE Level 2								
6.	SE Level 3								
7.	SE Level 4								
8.	SE Level 5								
9.	SE Level 6								
10.	SE Level 7	0.48							
11.	Bias Level 1	0.07	0.14						
12.	Bias Level 2	0.04	0.00	0.05					
13.	Bias Level 3	0.04	0.12	0.15	0.20				
14.	Bias Level 4	-0.17	-0.08	0.00	0.18	0.14			
15.	Bias Level 5	0.02	-0.06	0.04	0.06	0.21	0.26		
16.	Bias Level 6	0.17	0.07	0.07	0.07	0.13	0.13	0.17	
17.	Bias Level 7	0.03	-0.07	0.08	0.14	0.22	0.11	0.23	0.21
18.	Perf. Level 1	0.32	0.26	-0.21	-0.16	-0.08	-0.13	-0.23	-0.17
19.	Perf. Level 2	0.19	0.21	0.10	-0.10	-0.25	-0.14	-0.21	-0.09
20.	Perf. Level 3	0.24	0.33	-0.06	-0.25	-0.19	-0.15	-0.28	-0.35
21.	Perf. Level 4	0.18	0.29	-0.08	-0.17	-0.14	-0.25	-0.27	-0.12
22.	Perf. Level 5	0.24	0.25	-0.14	-0.17	-0.17	-0.21	-0.14	-0.15
23.	Perf. Level 6	0.21	0.29	0.00	0.05	0.02	-0.16	-0.26	-0.37
24.	Perf. Level 7	0.11	0.26	-0.08	-0.39	-0.02	-0.25	-0.13	-0.14

Correlations among Indicators in the Structural Equation Model (continued)

	17	18	19	20	21	22	23
1. HS Self							
2. HS GPA							
3. ACT Math							
4. SE Level 1							
5. SE Level 2							
6. SE Level 3							
7. SE Level 4							
8. SE Level 5							
9. SE Level 6							
10. SE Level 7							
11. Bias Level 1							
12. Bias Level 2							
13. Bias Level 3							
14. Bias Level 4							
15. Bias Level 5							
16. Bias Level 6							
17. Bias Level 7							
18. Perf. Level 1	-0.08						
19. Perf. Level 2	-0.12	0.23					
20. Perf. Level 3	-0.27	0.14	0.32				
21. Perf. Level 4	-0.29	0.27	0.47	0.35			
22. Perf. Level 5	-0.25	0.25	0.40	0.31	0.29		
23. Perf. Level 6	-0.05	0.28	0.21	0.41	0.35	0.38	
24. Perf. Level 7	-0.10	0.23	0.50	0.52	0.61	0.42	0.49

APPENDIX F

Final Code List for Qualitative Analysis of Interview Data

Code	Coded when Participant Referred to:	Frequency
High School Math	high school mathematics performance or achievement	7
College in High School	College-level calculus, college algebra, or statistics while in high school	13
Like/Dislike Teacher	Personal feelings about an instructor (separate from pedagogy)	10
Trauma	Strong negative reaction to the actions of an instructor.	3
Personality	Approachability, friendliness, “can to talk to”, funny, nice, etc.	9
Pedagogy	Pedagogical behavior of a math instructor	15
Work Harder	Perceived increased effort in a class because of feelings about an instructor	2
Sources of SE	Perceived reason for mathematical confidence	1
Physiological/Emotional	Emotions, fear, nervousness, anxiety	7
Social Persuasions	Comments from peers, instructors, friends, or family on math competency	12
Mastery Experiences	Results of attempts to solve mathematical problems, especially exams and homework	15
Good Performance	Perceived high math performance	8
Poor Performance	Perceived low math performance	18
Vicarious Experiences	Perceptions of others’ success or failure in mathematics, especially peers or family members	10
Why/When Became Math Teacher	Reasons for becoming a math teacher, reasons for choosing a grade band	13
Meaning of SE Scale	reasoning for choosing self-efficacy ratings on surveys	11
Afraid to put SE= 6	Aversion to the highest possible rating on the self-efficacy scale (6)	7

Optimism	Reported tendency to prefer higher self-efficacy scores because of personal optimism	2
Expect Minor Errors	Possibility of minor errors should not affect self-efficacy	6
SE = Understanding	Task- or content-specific evaluations of self-efficacy	5
Cut Off For Correct	Cut-off for SE ratings if asked to rate SE as YES or NO	4
Familiarity	specific experiences with content or similar tasks	15
Number of Steps	Evaluating self-efficacy based on the perceived number of steps needed to complete the problem (more steps = lower SE)	6
SE -> Performance	Direct belief that strong self-efficacy increased chances of success	2
Checking Out	Reduced effort or attendance based on dislike of a class or low self-efficacy	8
Retaking Classes	Experiences during second (or third) attempt at a class	6
Role of Teacher	Preference for new instructor when retaking a class	6
Math SE Trajectory	Perceived change in self-efficacy after completing a college math class	12
Failing Effect	Lowered SE after perceived low exam scores	8
Preview Effect	Increased SE after failing a math class (not an exam)	7
Math Identity	Self-beliefs about math skills, preferred learning style, personal work ethic in math.	17
Calibration	Alignment or misalignment between stated math SE and math performance	15