INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

MATHEMATICS SELF-EFFICACY CALIBRATION OF SEVENTH GRADERS

by

PEGGY PEI-I CHEN

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2002

UMI Number: 3047203

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for degree of Doctor of Philosophy.

4/25/2007
Date

Chair of Examining Committee

4/25/2007
Date

Executive Officer

Professor Shirley Feldmann

Professor Carol K. Tittle

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

MATHEMATICS SELF-EFFICACY CALIBRATION OF SEVENTH GRADERS

by

Peggy Pei-I Chen

Advisor: Professor Barry Zimmerman

This study investigated seventh graders' math self-efficacy calibration and its effect on students' math performance, individual differences, such as gender, as well as academic variables, such as previous math achievement, post-performance effort judgment, and post-performance self-evaluation. According to Bandura (1986), students' self-efficacy beliefs about their capability to perform affects how they make choices of activities, courses of action, amount of effort to exert, and length of time engaged on a task. To date, the issue of the accuracy judgment of self-efficacy beliefs, termed calibration, has received little investigation. In the present study, it was measured in two ways: accuracy, which measures the magnitude of judgment errors; and bias, which measures the direction of judgment errors. In addition, the design of the study enabled the researcher to assess the relationship between students' personal processes (e.g., self-efficacy judgments of math capability, calibration, effort judgment, and performance evaluation) and variations in context (e.g., specific math problems and their difficulty level).

The results revealed that students' calibration accuracy significantly increased the predictiveness of their self-efficacy beliefs. Path analysis showed that calibration

accuracy had both direct and indirect effects on math performance, with the indirect effects mediated through the students' self-efficacy beliefs. Self-efficacy played a direct role in predicting students' math performance, post-performance self-evaluation, and post-performance judgments of effort. The effects of prior math achievement on math performance were mediated largely through the students' self-efficacy beliefs. Unexpectedly, the effect of self-efficacy on post-performance judgments of effort was negative, indicating that high efficacy students needed to spend less effort in solving the math problems than low efficacy students. As for the individual differences in gender, the study found no statistical differences on any of the dependent measures, although boys had numerically higher self-efficacy, post-performance self-evaluation, and lower effort judgment than girls. In conclusion, the results revealed that students' self-efficacy beliefs play an important role in their acquisition of mathematical competence. Such information can be vital in assisting educators to tailor interventions that will enhance students' beliefs in their capability to learn math and as well as their actual success.

Acknowledgments

My gratitude to...
Professor Barry Zimmerman — the most excellent faculty advisor & mentor

My indebtedness to...
Professor Shirley Feldmann
Professor Carol K. Tittle
— the most dynamic committee members

My thanks to...
Professor John Hudesman
Professor Manual Martinez-Pons
— two most helpful readers

My appreciation to...
Gabby — the best editor and a good friend

My blessings to...
Fr. Steve
Theresa
— "lifelines" in Tennessee and data collection

My warm regards to...
Theresa, Ed, Claude,
Heather, Annie, and David
— the most gracious friends and support

My affection for...
Kiki, my cat — the most faithful companion of 13 years.

My **LOVE** to...
Dad and Mom — the most wonderful parents on earth

THANK YOU!!!

Table of Contents

Approval Page iii

Abstract iv

Acknowledgments iv

Table of Contents vii

List of Tables ix

List of Figures x

Chapter

I INTRODUCTION 1

II LITERATURE REVIEW 6

Overview of the Self-Efficacy Theory 7
Uniqueness of Self-Efficacy and Its Influence on Learning and Achievement 10
Research on Calibration 14
Mathematics and Gender 20
Research Studies on Effort and Self-Efficacy 32
Research Studies on Mathematics Self-Efficacy 35
Synthesis and Hypotheses 62
Objectives and Hypotheses 70

III METHODOLOGY 74

Participants 74
Instruments 74
Math Performance Test 74
Mathematics Self-efficacy Measure 78
Academic Motivation Measure—Effort Judgment 79
Self-evaluation Measure 79
Previous Math Achievement 80

Procedures 80
Data Analysis 83

IV RESULTS 87

Reliability of Test Forms 87
Difficulty Level of the Test Items 87
Role of Gender and Item Difficulty on Dependent Measures 88
Influence of Calibration Accuracy and Item Difficulty Levels on Self-efficacy, Post-performance Self-evaluation, Effort Judgment, and Math Performance 92
Correlations 98
Regression Analysis 99
Path Analysis 99

V DISCUSSION 106

General Discussion 106
Hypotheses 107
Hypothesis 1 107
Hypothesis 2 110
Hypothesis 3 113
Hypothesis 4 115
Hypothesis 5 116
Hypothesis 6 117
Summary 119
Educational Implications 120
Directions for Future Research 121
Conclusion 122

APPENDICES

A Form 1 of Student Packet 123

B Form 2 of Student Packet 142

REFERENCES 161

LIST OF TABLES

Table

1 Means and Standard Deviations for Calibration Accuracy, Bias, Self-efficacy, Effort Judgment, and Self-evaluation as a Function of Gender and Level of Item Difficulty 89

2 Post-hoc Analysis on Dependent Measures of Calibration Accuracy, Bias, Self-efficacy, Effort Judgment, and Self-evaluation as a Function of Level of Item Difficulty 91

3 Correlations among Measures 98

4 Decomposition of Effects from the Path Analysis 103

5 Direct, Indirect, and Total Effects on Math Performance, Self-evaluation, Effort Judgment, and Self-efficacy 105

LIST OF FIGURES

Figure

1 Means of Self-efficacy of High- and Low-Calibrated Students according to Item Difficulty Levels 94

2 Means of Post-performance Self-evaluation of High- and Low-Calibrated Students according to Item Difficulty Levels 96

3 Means of Math Performance of High- and Low-Calibrated Students according to Item Difficulty Levels 97

4 Proposed Path Model 100

5 Final Reduced Path Model 102

Chapter I

INTRODUCTION

Perceived academic self-efficacy has been one of the most studied constructs in the field of academic motivation research. It is unique from other related motivational constructs such as self-concept because perceived self-efficacy is specific and closely corresponds to target behaviors or performance. Research has shown that perceived academic self-efficacy positively influences student academic choices, academic performance, effort and persistence (Bong & Clark, 1999). The role of self-efficacy in students' math performance has also been studied. For example, Pajares and Kranzler (1995) used path analysis to examine the effect of mathematics self-efficacy and general mental ability on the math problem-solving performance of high school students. They found that math self-efficacy and general ability each had strong direct effects on math performance. There is also evidence that math self-efficacy predicts mathematics performance for middle school students (Pajares & Graham, 1999). Further, math self-efficacy predicts math performance better than general self-concept measures, indicating the importance of its greater specificity. Thus, the convergent and discriminant validity of perceived self-efficacy has been established, and perceived self-efficacy uniquely predicts students' achievement and motivation to learn (Zimmerman, 2000).

Bandura (1997) stated that the self-efficacy construct has three dimensions: level, strength, and generality. Most research on self-efficacy has focused on self-efficacy levels and strength dimensions (Ewers & Wood, 1993; Pajares & Miller, 1994; Pajares &

Valiante, 1999). Recent research focused on the dimension of generality of academic self-efficacy among high school students. Using factor analysis, Bong (1997b) examined the degree of generality among perceptions of academic self-efficacy for different subject areas such as English, Spanish, U. S. history, algebra, geometry, and chemistry. Although task-focused measures of self-efficacy have been found to be the best predictors, the researcher found evidence that self-beliefs generalize to similar skills (Bong, 1997b).

Although the three dimensions of self-efficacy delineated by Bandura (1977) have been well researched, the issue of the accuracy of students' self-perceptions in relation to their actual performance, or calibration, has only recently surfaced in the literature (Ewers & Wood, 1993; Pajares & Graham, 1999; Pajares & Kranzler, 1995). Bandura (1986) stated that reasonably matched self-efficacy judgments and actions are most desirable, even though higher self-efficacy judgments can enhance motivation to improve future performance. However, recent studies have shown that many students perceived high self-efficacy but their performance was not in accordance with this perception; in other words, they were overconfident when making judgments about their performance (Ewers & Wood, 1993; Pajares & Graham, 1999; Pajares & Kranzler, 1995). Ewers and Wood (1993) studied fifth grade gifted and regular students' math self-efficacy by examining self-efficacy level, self-efficacy strength, and prediction accuracy. They found that gifted students made fewer overestimations than regular students, and boys made more overestimations than girls. Pajares and Kranzler (1995) studied high school students' math self-efficacy and their general ability in math problem-solving. Most

students of either gender were found to be overconfident in their math capabilities. In another study of middle school students, Pajares and Graham (1999) again found that (1) students were overconfident in their self-efficacy judgments, (2) gifted students had less overconfidence than regular students, and (3) there were no gender differences in calibration.

Although a few researchers have examined the issue of self-efficacy calibration, the approach of assessing calibration has its limitations. For example, in Pajares and Graham's (1999) study, students' math performance scores (dependent variables) were used as one of the measures of calibration (independent variables) and this raises the possibility of confounding. Further, calibration of self-efficacy has been measured differently by different researchers (Ewers & Wood, 1993; Pajares & Graham, 1999; Pajares & Miller, 1997). Thus, the question of how to assess self-efficacy calibration remains unanswered.

The present research sought to address the measurement issues of self-efficacy calibration. An important question to answer was whether calibration is another dimension of self-efficacy construct that contributes to its predictive power, as mentioned by Bandura (1997). In addition, the question pertaining to sources of individual differences in calibration was tested. According to Ewers and Wood (1993), boys were found to be more overconfident than girls, but Pajares and Kranzler (1995) did not obtain the same results. Therefore, it is important to assess student gender to further understand how perceived self-efficacy and its calibration differ and contribute to math achievement.

The present research also sought to determine what influences students' self-efficacy calibration. An important question to explore was whether students' judgments of the difficulty levels of a task influence their calibration of self-efficacy. Do students misjudge the difficulty levels of their tasks, thus misperceiving their capabilities to solve the tasks? Another question was: Why do well-calibrated students have higher math performance than poorly-calibrated students? One possible explanation is that the self-efficacy of well-calibrated students remains more stable as they actually begin to solve math problems. These students are probably not surprised by the level of task difficulty they encounter, but poorly-calibrated students are surprised, thereby producing subsequent declines in self-efficacy. The inability of poorly-calibrated students to judge task difficulty remains an unstudied topic to date.

In summary, the present research attempted to address the following objectives:

1. To determine whether calibration measures of mathematics self-efficacy are an additional dimension of self-efficacy.
2. To investigate the separate and combined influence of mathematics self-efficacy, math self-efficacy calibration on motivation factors (effort judgment and self-evaluation), and math performance.
3. To examine possible gender differences on math self-efficacy strength, calibration, motivation factors, and math performance.
4. To test whether students' previous math achievements influence their math self-efficacy, math self-efficacy calibration, motivation factors, and math performance.

5. To study students' self-evaluation and effort judgments of their math performance after solving math problems.
6. To investigate students' math self-efficacy and math self-efficacy calibration in relation to levels of task difficulty (e.g., easy, moderate, and difficult).

Chapter II

LITERATURE REVIEW

The literature review consists of seven sections. The first section provides an overview of Bandura's self-efficacy theory. In this section, the dimensions and measurement issues of self-efficacy are delineated. The second section, based on Bandura's self-efficacy theory, discusses its uniqueness as a motivational construct and its influence on academic learning and achievement. In this section, subtle differences of the self-efficacy construct and its influence on academic motivation and achievement are documented. Also in this section, evidence of the self-efficacy construct in relation to achievement and motivation is quantitatively shown through meta-analysis. In the third section, research relating to calibration and accuracy judgments is discussed. Specifically, differences between measuring accuracy and association as well as measurements of calibration are examined. In the fourth section, the discussion focuses on mathematics and gender differences. Specifically, gender differences in mathematics achievement, problem solving, attitudes, and affect are scrutinized. The fifth section contains studies on a motivational variable, effort, and its relationship to self-efficacy. Specifically, the section examines the positive relationship between self-efficacy and effort and their contribution in predicting performance. The sixth section interweaves Bandura's self-efficacy theory and the topics discussed earlier: understanding self-efficacy as a unique construct; examining issues of calibration; and documenting gender differences in mathematics. In addition, specific research studies on mathematics self-

efficacy and self-efficacy calibration in relation to gender and academic performance are discussed. The seventh section presents a synthesis of research studies examined in this literature review as a foundation for the hypotheses presented in this study.

Overview of the Self-Efficacy Theory

In this section, self-efficacy is defined and explained. Next, three dimensions of self-efficacy are discussed and delineated. Finally, measurements of each self-efficacy dimension are described with relevant examples.

According to Bandura (1997), self-efficacy refers to an individual's judgments of his or her capabilities to perform tasks at designated levels. One's self-efficacy can affect his or her choice of activities and environmental settings. People choose to engage in activities in which they believe they can succeed. Self-efficacy also affects the amount of effort and length of time people will persist when they encounter challenging circumstances. For example, a person with high self-efficacy tends to put forth much effort on difficult tasks, works hard, and persists to overcome obstacles. By comparison, a person with low self-efficacy expresses doubts about his or her capability and is unwilling to expend effort and time on difficult tasks.

Self-efficacy is a multidimensional construct (Bandura, 1997). The structure of self-efficacy consists of level, strength, and generality. Level refers to one's perceived self-efficacy of the task demands within a particular domain. Since self-efficacy beliefs are context-driven, a person may not feel competent to successfully complete a task if he or she judges the task to be too demanding. In other words, contextual circumstances are

the requirements of one's performance by which a person makes perceived efficacy judgments. Strength refers to the perceptions of the amount that one "can do" on a particular task. The stronger the perceived self-efficacy, the greater the chance of successfully executing the chosen task. However, Bandura (1997) also asserted that the strength of perceived self-efficacy was not always linear in relation to one's chosen behavior. "A certain threshold of self-assurance is needed to attempt a course of action, but higher strengths of self-efficacy will result in the same attempt" (p. 43). Generality refers to one's transferable perceived self-efficacy across a wide range of domains or among levels of functioning on a particular activity. Generality can vary on a number of dimensions which include: the degree of task similarity; the forms in which competence is expressed such as behavioral, cognitive, or affective; the qualitative features of tasks; and the characteristics of the actors toward whom the behavior is directed (Bandura, 1997).

Bandura (1997) argued that when measuring perceived self-efficacy, the most appropriate phrase to be used should be "can do" rather than "will do." He further posited that the tasks for which people are asked to make self-efficacy judgments should be constructed with a sufficient level of challenge in order to avoid ceiling effects. According to Bandura, self-efficacy items are generally measured on a Likert-type scale, ranging from 0 to 100 with intervals of 10. For example, people record their strength of beliefs to complete a task on a scale ranging from 0 ("cannot do"), through 50 ("moderately certain can do"), to 100 ("certain can do"). Similar measurement structures can be used with single-digit ranges and a unit of one interval. However, Bandura (1997)

cautioned researchers that a self-efficacy scale should not use short ranges with few intervals because it is not sensitive enough to capture people's efficacy judgments. The format of the scale is usually the same throughout the tasks being assessed. The self-efficacy strength scores are then summed across total items and divided by the total number of items. The measure of efficacy level can be extracted by selecting a cutoff value, below which people would judge themselves incapable of successfully executing the tasks in question. For example, Wood and Ewers (1993) obtained students' math self-efficacy levels by counting the number of items with a self-efficacy rating of above 3 (on a 5-point scale with 1 "not at all sure" to 5 "totally sure").

Assessments of the generality dimension of self-efficacy have not been prevalent. To better measure the generality of self-efficacy, Bong (1997b) used factor analysis. She examined the degree of generality among perceptions of academic self-efficacy for different subject areas such as English, Spanish, U. S. history, algebra, geometry, and chemistry. Although task-focused measures of self-efficacy have been found to be the best predictors, evidence indicates that self-beliefs generalize to similar areas of skills. Bong (1997b) found that students' perceptions of self-efficacy extended beyond the boundary of a specific problem to other problems within a subject area. There is also evidence that self-efficacy for each school subject formed a first-order factor.

In summary, Bandura (1997) described self-efficacy as the personal judgment of one's capability to execute a course of action in order to complete a task successfully. Self-efficacy consisted of three dimensions: level, strength, and generality. Much of the research on self-efficacy has focused on its strength and level. Only recently, Bong

(1997) used factor analysis to test the generality dimension of academic self-efficacy across subject domains and found that the generality of academic self-efficacy was partly dependent on similarity among tasks.

Bandura (1997) cautioned researchers about the measurement issues of self-efficacy. He posited that self-efficacy should be measured in correspondence to the tasks in question and that the tasks themselves should comprise various levels of difficulty. Scales used to assess self-efficacy should contain a wide range of intervals, thus sensitively capturing people's judgments.

Uniqueness of Self-Efficacy and Its Influence on Learning and Achievement

Based on Bandura's theory of self-efficacy, Zimmerman (2000) highlighted various research studies on the construct, its influence on student learning, and its interaction with academic self-regulation processes. Again, self-efficacy pertains to an individual's judgments of his or her capability to organize and perform tasks in order to meet designated goals. Zimmerman (2000) discussed a few unique components of the self-efficacy construct. Self-efficacy is a measure of one's performance capabilities, not personal traits or characteristics. It is multidimensional, yet domain-specific. An example of this is when a student's efficacy beliefs about performing on a math task differ from those for a writing task. Self-efficacy measures are sensitive to various performance contexts; thus, performing in a noisy place may be different from learning in a quiet place. Self-efficacy measures depend on mastery criteria rather than normative criteria. One would be asked to make judgments about his or her capability to

successfully complete an activity rather than compare his or her capability and success to others. Self-efficacy judgments pertain to future functioning and should be measured before performing the target tasks.

Zimmerman (2000) differentiated a self-efficacy construct from closely related constructs such as outcome expectancies, self-concept, and perceived control. He posited that self-efficacy is different from outcome expectancies because it pertains to a perceived competence of performing tasks; on the other hand, outcome expectancies refer to attaining certain outcomes. Self-efficacy also differs from self-concept in its focus on specific tasks and corresponds to the performance in question; in turn, self-concept focuses on a variety of personal characteristics that can be general in scope. Lastly, Zimmerman differentiated between self-efficacy and perceived control. Measures of perceived control focus on general beliefs about whether outcomes are contributed by the actor or by external forces. Zimmerman cited evidence from various studies which showed self-efficacy to be a better predictor of academic outcomes than constructs such as outcome expectancy (Shell, Murphy, & Bruning, 1989), self-concept (Pajares & Miller, 1994), and perceived control (Smith, 1989).

Zimmerman (2000) also discussed the role of self-efficacy in relation to various motivational variables such as persistence, choice of activities, effort, and emotional reactions. Self-efficacy beliefs were shown to influence academic motivation and learning. Research showed that students with a higher sense of efficacy were more likely to choose challenging activities, exert more effort when encountering difficulties, increase persistence when engaging activities, and experience less academic anxiety.

Self-efficacy beliefs further influence students to self-regulate their learning process. Zimmerman posited that self-efficacy beliefs influence students' motivation for learning by self-regulating their learning processes, including setting goals, monitoring and evaluating learning strategies, and using and modifying effective learning strategies. Based on this assertion, a high sense of self-efficacy promotes greater motivation to learn, and students become engaged in self-regulated learning that, in turn, produces higher academic achievement. Zimmerman provided a summary of research on the self-efficacy construct and its influence on student motivation and learning. Further, through various research studies, the convergent and discriminant validity of perceived self-efficacy was established, and perceived self-efficacy was found to uniquely predict students' achievement and motivation to learn.

To provide further evidence of the validity of the self-efficacy construct in relation to academic achievement and motivation, Multon, Brown, and Lent (1991) used meta-analysis to show the relation between self-efficacy beliefs and academic performance and persistence. A meta-analysis is a statistical technique to quantify evidence from a number of studies (Hedges & Olkin, 1985). The researchers hypothesized that self-efficacy beliefs were positively related to academic performance and academic persistence. In order to select appropriate studies to include in meta-analysis, three criteria were implemented: (1) a measure of self-efficacy, (2) a measure of performance and/or persistence, and (3) sufficient information to calculate effect size. Studies published between 1988 to 1997 were considered for the study. A total of 39

studies were included; 38 of these studies were used for meta-analysis of performance and 18 of these 39 studies were used for meta-analysis of persistence.

For self-efficacy and performance meta-analysis, the results showed an overall effect size of .38. Across various types of student samples, designs, and criterion measures, self-efficacy accounted for 14% of the variance in students' academic performance. The researchers further examined variables that mediate effect size based on performance-efficacy meta-analysis. Four sources of effect size variance were found: (1) time of assessment, (2) students' achievement level, (3) participants' age, and (4) types of performance measure. The researchers found that effect sizes of self-efficacy and performance were stronger when estimated from posttreatment ($\underline{d}$ = .58) rather than from pretreatment ($\underline{d}$ = .32). One interpretation was that this effect size difference was a result of the self-efficacy-inducing treatments used in experimental studies, not only changing efficacy beliefs but also strengthening efficacy-performance relationships. The researchers also found that the relationship between efficacy and performance varied according to students' achievement levels, with higher effect size for lower-achieving students ($\underline{d}$ = .56) and smaller effect size for average-achieving students ($\underline{d}$ = .33). As for the influence of the age of students on performance-efficacy, high school and college-age students displayed larger effect sizes ($\underline{d}$ = .41 and $\underline{d}$ = .35, respectively) than elementary school-age students ($\underline{d}$ = .21). Finally, measures such as basic skills produced a stronger effect size ($\underline{d}$ = .52), followed by classroom-based performance ($\underline{d}$ = .36) and standardized achievement tests ($\underline{d}$ = .13).

For efficacy and persistence meta-analysis, Multon, Brown, and Lent (1991) found an overall effect size of .34, indicating that self-efficacy accounted for 12% of persistence. Interestingly, effect sizes for self-efficacy and persistence differed greatly, depending on how persistence was operationalized. When persistence was a measure of time spent on tasks, the effect size was .17, compared to a measure based on number of items/tasks completed or attempted ($\underline{d}$ = .48). Their meta-analyses showed that self-efficacy beliefs were significantly and positively related to academic performance and persistence on tasks. Although effect sizes could vary according to different experimental designs, time of self-efficacy and performance, and participant age and achievement levels, this study quantitatively demonstrated the unique contribution of self-efficacy in academic achievement and motivation.

In summary, the discussion of this section first focused on the uniqueness of the self-efficacy construct compared to other motivational constructs, and its influence on students' academic achievement and motivation (Zimmerman, 2000). The discussion then sharpened its emphasis on the validity of the self-efficacy construct. By using a quantitative method, meta-analysis, Multon, Brown, and Lent (1991) found self-efficacy to have a significant and positive influence on academic achievement and motivation.

Research on Calibration

Thus far, the discussion has attempted to define, explain, differentiate, and validate the self-efficacy construct, both qualitatively and quantitatively. To understand self-efficacy calibration and its influence on students' academic achievement, a closer

look at calibration and its measurement issues is warranted. Schraw (1995) discussed the different measures of feeling-of-knowing accuracy often reported in research. He first distinguished two ways to measure feeling-of-knowing: accuracy and association. Agreement accuracy was defined as "the degree to which feeling-of-knowing judgments and observed events agree," while association referred to "the degree to which levels of one variable covary with levels of a second variables" (pp. 322-323). Schraw examined two measures: (1) the Hamann coefficient (HC) which measures agreement accuracy, and (2) Goodman and Kruskal's gamma (γ) which measures association. The researcher showed mathematical derivations and examples to distinguish between agreement accuracy and association. Even though the Hamann coefficient (HC) and gamma (γ) are mathematically similar, they do differ in two ways: (1) observable range value (HC from -1 to 1; γ from 0 to 1) and (2) HC sensitivity to the number of incorrect responses (e.g., misclassification) in feeling-of-knowing. Schraw asserted that Goodman and Kruskal's gamma (γ) is not appropriate for measuring agreement accuracy. Although the Hamann coefficient is suitable for measuring accuracy, it is problematic because "it is not generalizable to a data array of unequal dimension (e.g., 2 x 3)" (p. 329).

As an alternative to the Hamann coefficient, Schraw (1995) suggested bias and accuracy measures. These two measures of agreement accuracy were based on the Euclidean distance between predicted and observed performance. Bias refers to the measure of "the direction of judgment errors" while accuracy refers to the measure of "the magnitude of judgment error" (p. 329). Bias is computed by taking the mean differences between predicted and observed scores that range from -1 to 1, with scores

greater than 0 indicating overconfidence and less than 0 underconfidence. Accuracy can be computed in two ways: by squaring the bias scores or taking the absolute value of bias scores; accuracy scores range from 0 to 1. The following empirical study employed measures of bias and accuracy, along with a coefficient alpha and a point-biserial, to assess calibration.

Schraw, Potenza, and Nebelsick-Gullet (1993) studied calibration of academic performance using an experimental design. They were interested in understanding how student calibration improved under different experimental manipulations, specifically whether feedback and external incentives affected test performance and calibration. One rationale they gave was that providing feedback during testing could improve calibration for subsequent tests. Another rationale was that one possible way to improve calibration was providing learners with external incentives such as rewards, especially since students may be unmotivated to engage in experimental tasks.

Schraw et al. (1993) used four different measures to assess calibration. First, mean bias was used to measure the direction of judgment errors. Second, mean accuracy was used to measure the magnitude of judgment error. Third, a coefficient alpha was used to measure the internal consistency of performance scores. For example, an alpha > .70 suggested that students' judgments about their performances were consistent, regardless of item difficulty or actual performance. Fourth, a point-biserial correlation was made between judged performance and actual performance. Mean bias and mean accuracy scores were compared across both easy and difficult tests. On easy tests, 80% of the questions were answered correctly, while on difficult tests, only 50% of the

questions were answered correctly. The researchers thought that by comparing bias and accuracy scores across tests with various difficulty levels, they could assess whether possible effects on these measures were due to test difficulty level.

Schraw et al. (1993) posed four hypotheses in their study. The first hypothesis was that easy rather than difficult questions would lead to less overconfidence and better calibration. Second, the researchers hypothesized that the incentive to improve calibration would decrease mean bias and increase calibration accuracy. By giving the students incentives to improve calibration, they were encouraged to reflect on their comprehension and performance (promoted self-generated cognitive feedback during testing); thus, this would, in turn, decrease mean bias and increase calibration accuracy. The third hypothesis was that the coefficient alpha would be high (e.g., $\alpha > .90$) due to participants' subjective estimations of their performance versus objective information such as test item difficulty. Fourth, the researchers expected that correlations would be higher in the incentive groups than in the control group.

Eighty-five undergraduate students participated in the study, which was a 2 x 3 x 2 experimental design with two types of feedback (feedback vs. no feedback), three types of incentives (extra credits for improved performance, extra credits for calibration, and control), and two types of tests (easy and difficult). The easy test contained 36 multiple-choice questions in eight categories regarding reading comprehension (Nelson-Denny Reading Comprehension); the difficult test consisted of eight math questions (e.g., computing simple probabilities and combinatorial problems). The difficulty levels of these tests were based on a pilot study and normed on 10 students per test. The overall

performance score was 76% for the easy test and 46% for the difficult test. With each test question came a rating scale along which students made their perceived accuracy judgments. This procedure was a post-hoc procedure: because the accuracy judgment was made after solving each question. The confidence scale ranged from no accuracy (0%) to perfect accuracy (100%). For feedback conditions, one group of students received no feedback, while the experimental group received correct solutions after taking each subtest as a form of feedback. For the incentive condition, one group of students was told they would receive double credits for improving performance if their scores were above the group mean. Another incentive group of students was told they would receive double credits if they improved their calibration to within 25% of a perfect calibration. The third incentive group was a control group.

The results showed main effect incentives and tests. The researchers found that students who received extra credits (incentives) for improving calibration performed better than the control group. Post-hoc tests showed that the incentive-calibration group had higher confidence scores than the control group. Further, this group, which received incentives for improving calibration, made more accurate judgments about their performances. As for the test variable, the easy test (reading comprehension) was associated with higher performance, greater and lesser overconfidence, and greater accuracy. The researchers found a very high internal consistency on judged performances ($\alpha = .92$), but only a moderately high internal consistency on actual performance ($\alpha = .69$). In other words, students were more consistent about their judged performance than their actual performance on individual items. Finally, the researchers

used point-biserial correlations between judged and actual performance scores across easy and difficult tests for each student. Contrary to the researchers' hypotheses, these six experimental groups did not differ from one another across levels of test difficulty. The researchers concluded that the point-biserial nonparametric correlations were insensitive measures of differences among the groups.

Results of this study supported the hypothesis that incentives for calibration do affect performance and calibration. Students who received external rewards to improve calibration performed better, were less biased, and were more accurate than students who received rewards for improving performance. As for feedback, this study found no effect on performance, bias, or accuracy scores. The researchers concluded that feedback was of little value to learners because students self-generated feedback during the testing. The researchers theorized that incentives were more effective than feedback in improving calibration because of their motivational importance. The hypothesis that students were more accurate when calibrating their performance on easy rather than difficult questions was also supported. However, this study has its limitations. First, the self-judgment measure was post-hoc (i.e., students made their accuracy judgments after performing on questions) unlike self-efficacy measures. The process of answering in each question can provide self-evaluation opportunities and performance feedback. Further, the easy and difficult tests came from very different subject matter areas (reading comprehension vs. math); thus, the main effect of the tests may be due to the subject matter rather than to the level of test difficulty.

In summary, research on calibration provided several measures of agreement accuracy and association. It is crucial to differentiate, qualitatively and quantitatively, between these two classes of measures. Calibration was best measured in terms of bias and accuracy (Schraw, 1995). These measures (bias and accuracy) were found to be sensitive to variations in external rewards (e.g., received extra credits for improvement).

Mathematics and Gender

Mathematics in relation to gender has been extensively investigated from the perspectives of career options, occupational interest, and educational choice. Various theoretical models have been developed to explain any differences in mathematics existing between males and females. In the first part of this section, the importance of mathematics performance, career interests in mathematics, and attitudes and affects toward mathematics in relation to gender is discussed (Eccles, 1994; Hyde, Fennema, & Lamon, 1990; Hyde et al., 1990). The second part of this section focuses on one specific study regarding gender and cognitive and motivational behaviors during mathematics problem-solving (Vermeer, Boekaerts, & Seegers, 2000).

Eccles and her colleagues have studied gender in relation to educational and occupational choices for a number of years. More specifically, they have researched the motivational and social factors that influence participants' achievement goals, career interests and choices, course selections, persistence and effort across achievement-related activities. Eccles (1994) summarized her research in this area over the past 15 years to

answer an important question: "Why do women choose the particular occupations they do?" (p. 604).

Eccles' (1994) achievement-related model was developed from several theories such as decision-making models, achievement theories, and attribution theories. Eccles' model explained achievement-related choices made directly from beliefs in (1) expectation of success and (2) subjective task values (e.g., incentive and attainment value, utility value, and cost). Her model depicted a number of variables that influenced these beliefs, such as cultural milieu; socializers' (parents and teachers) beliefs, aptitudes, prior achievement experiences; and children's interpretations and perceptions of the culture, socializers' beliefs, aptitudes, and experiences. Eccles (1994) hypothesized that people who make achievement-related choices are often guided by "(a) one's expectations for success in, and sense of personal efficacy for, the various options; (b) the relation of the options both to one's short- and long-range goals and to one's core self-identity and basic psychological needs; (c) the individual's gender role schema; and (d) the potential cost of investing time in one activity rather than another" (pp. 591-592). Eccles and her colleagues tested many hypotheses delineated in this model.

Eccles (1994) discussed a number of empirical studies to better understand why women choose the particular careers they do. Using her expectancy-value model, she was able to demonstrate that gender differences are linked to differences in individual expectations for success and subjective task values regarding educational and occupational choices. She concluded that females are less likely than males to choose fields relating to mathematics and science because females have less confidence in their

abilities and place less value on these fields than on other occupational fields. Eccles further suggested that gender socialization at home, school, and among peers influenced and shaped students' self-perceptions and task values. She sought to explain gender differences in relation to educational and occupational achievement by evaluating individual studies according to her model. This comprehensive model included such variables as culture milieu, child's aptitudes, child's belief system, child's interpretation of experiences, and child's affective memories. Thus, Eccles' (1994) expectancy-value model provided considerable guideline for studying cognitive and motivational issues underlying gender and mathematics.

To test gender differences in mathematics performance, Hyde, Fennema, and Lamon (1990) conducted a meta-analysis. The investigators were interested in the following issues: (1) the magnitude of gender difference in mathematics performance; (2) the magnitude of gender difference relating to the cognitive level of the tasks (e.g., computational, conceptual, problem solving or mixed); (3) the magnitude of gender difference relating to math content (e.g., arithmetic, algebra, geometry, calculus or mixed); (4) the developmental (age) gender difference; (5) the magnitude of gender difference in math relating to ethnicity; (6) the magnitude of gender difference in mathematics performance across various samples (e.g., national sample, general sample, highly selective sample, etc.); and (7) increase or decrease of gender difference in mathematics performance over the years.

One hundred studies were included for the meta-analysis, yielding 254 independent effect sizes. Positive effect sizes indicated a favoring of male participants

and negative effect sizes a favoring of females. Overall, 51% of the effect sizes were positive, 6% were zero, and 43% were negative. However, overall, the effect sizes were not homogeneous; thus, the researchers partitioned the effect sizes into more homogeneous subsets. In relation to gender difference in mathematics performance, the researchers found a small positive mean effect size (d = .15). However, when the researchers looked at only general population samples, they found a small negative mean effect size (d = -.05). Females in the general population outperformed males by a very small number.

As for gender difference in mathematics in relation to cognitive level, Hyde, Fennema, and Lamon (1990) found very small effect sizes. They found females to be superior in computation (d = -.14) and males to be superior in problem solving (d = .08); they found no gender difference in understanding concepts (d = -.03). When mathematics test content was examined, no gender differences in arithmetic and algebra performance were noted. However, males performed slightly better in geometry (d = .13) and mixed math (d = .20). Interestingly, age difference showed only a small superiority in math performance among females during elementary (d = -.06) and middle (d = -.07) schools. However, there was a larger male superiority in math performance during high school (d = .29) and college (d = .41). As for ethnicity, the results showed no to very slight gender difference in math performance for Black (d = -.02), Hispanic (d = 0), and Asian (d = -.09) Americans. For White Americans, there was a small superiority of males in math performance (d = .13). The results also showed that gender difference was larger for highly selective samples (d = .54) and extremely selective samples (d = .41),

showing male superiority in math performance. When studies were divided into two subsets based on publication date, the results showed that studies published before 1973 reported a larger mean effect size favoring males (d = .31), compared with the mean effect size for studies published after 1973 (d = .14).

Hyde, Fennema, and Lamon (1990) conducted additional regression analyses to test the sources of variance in effect size. They found that 87% of the variance in effect size was explained by subject's age, selectivity of the sample, and cognitive level of the test. Of the three variables, age was the strongest predictor of effect size, followed by selectivity of the sample and cognitive level. In conclusion, this meta-analysis showed that overall gender difference in math performance was small. However, when looking at effect sizes within specific educational and task contexts, gender differences can be found. For example, during high school and even more so in college, females began to perform less successfully than males. Further, females performed less well on mathematics problem-solving tasks than males. These findings are critical because it is important to sustain female mathematics performance when girls move to high school and college. More important, problem solving is critical for success in many math-related and science-related educational and occupational fields.

In another meta-analysis, Hyde et al. (1990) compared gender in mathematics attitudes and affect. The study attempted to explore the following questions: (1) magnitude of gender differences in mathematics attitudes and affect; (2) developmental and gender differences on mathematics attitudes and affect; and (3) trend of gender differences in mathematics attitudes and affect over time. Measures of math attitudes and

affect constructs were based on Fennema and Sherman's (1976) scales and the Mathematics Anxiety Rating Scale (Richardson & Suinn, 1972). Fennema and Sherman's scales measured confidence in learning mathematics, mathematics anxiety, usefulness in mathematics, beliefs about math as a male domain, attitude toward success in mathematics, effectance motivation in mathematics, parental attitude reported by students, and teacher's attitude. The Mathematics Anxiety Rating Scale, assessing math anxiety in everyday and academic situations, was also examined.

In Hyde et al.'s (1990) meta-analysis, 70 articles were examined, yielding 126 effect sizes. Similar to Hyde, Fennema, and Lamon's (1990) study, positive effect sizes indicated higher scores for males while negative effect sizes indicated higher scores for females. The results showed that all effect sizes were small with one exception: the stereotyping of math as a male domain ($\underline{d}$ = -.90). This finding indicated that more males than females believed math to be a male domain. When examining age and gender differences in math attitudes and affect, the researchers found larger effect sizes for high school students' perceptions of mother's attitudes, father's attitudes, and teacher's attitudes, compared with other age groups. The age trend for attitudes toward mathematics was a positive increase of effect sizes, which indicated an increase in gender differences (males had more positive attitudes as their age increased). There was an overall positive effect size when examining mathematics self-concept or confidence (d = .16), with the largest effect size in the high school-aged group (d = .26). As for math anxiety, gender differences again favored males, indicating a lower math anxiety among males than among females. Hyde et al. (1990) conducted regression analyses to

determine sources that predicted effect sizes; these predictors included age, selectivity of sample, and year of publication of studies. Again, gender difference in mathematics attitudes and affect was larger with the high school-age group than the other groups. This meta-analysis showed that gender differences in mathematics attitudes and affect were generally small.

Effect sizes that showed gender differences, however small, indicated a pattern of females with more negative attitudes. With the two meta-analyses (Hyde, Fennema, & Lamon, 1990; Hyde et al., 1990) of gender differences in mathematics performance and attitudes and affect, it was interesting to note that the high school-age group of students showed the most gender differences. Further examination is necessary to explain why the transition from middle school to high school contributed to such changes.

A study by Vermeer, Boekaerts, and Seegers (2000) sought to understand middle school students' mathematics cognitive processes and affect beliefs. These researchers in the Netherlands studied the mathematical problem-solving behavior of sixth graders in relation to mathematical computation and application. Specifically, the researchers were interested in male and female student differences in cognitive and affective changes before, during, and after tasks. They measured task-specific variables, students' appraisals, and learning intentions directly before and after task execution. They also measured behavior-related variables, student performance, persistence following failure, and students' confidence/doubt during the problem-solving process. The first hypothesis of the study was that students were expected to have more positive appraisals and higher learning intentions with computation than with application problems. The second

hypothesis was that students were likely to express higher confidence judgments and persistence after failing a computation problem rather than an application problem. The third hypothesis was the existence of possible gender differences for variables measured with application problems, but not with computation problems. The researchers predicted that, compared to boys, girls were likely to express negative appraisals, show less confidence while solving application problems, and give up easily after solving an application problem incorrectly.

Using stratified random sampling, Vermeer et al. (2000) selected 160 sixth graders (80 girls, 80 boys) from 12 schools. After two students left the schools during the research period, 158 (79 girls, 79 boys) students remained in the study; they were predominantly Caucasian and from middle-class families. In the study, the researchers utilized five measures. One measure assessed students' abstract reasoning ability that consisted of analogies and categories subscales. The second measure, the On-Line Motivation Questionnaire (OMQ), assessed students' task-specific appraisals, learning intentions, and attributions. This scale consisted of two parts: one part was administered before students began the task, while the second part was given after the completion of the task. The first part consisted of appraisal questions (e.g., expectations, self-efficacy judgments, and task-difficulty perceptions) and learning intention questions measuring students' commitment to the task (e.g., "How much effort are you going to put into this task?"). The second part of the scale consisted of questions that assessed students' attributions. Students were asked to judge how well they had done with the task and what causes (capability, pleasure, luck, effort, and difficulty level) were attributed to the

results. The third measure assessed students' task performance on six computation and application question pairs. Vermeer, Boekaerts, and Seegers (2000) constructed these questions in pairs so that problems could be solved by using the same math skills for each pair. Because the researchers were interested in how students persisted after a failing experience, they asked students to solve one difficult problem after another. In other words, the first two math problems could be solved by 50% to 75% of same-age students, followed by two difficult questions that could be solved only by 50% of same-age students. The fourth measure was the Confidence and Doubt Questionnaire (CDQ), which assessed students' confidence level during math problem solving. The questionnaire consisted of drawings of five faces ranging from very sad (very doubtful) to very happy (very confident) expressions. During each problem-solving effort, the CDQ was presented to students several times on the test paper in order for them to make confidence judgments. The fifth measure assessed persistence, particularly related to students' reactions after failing to solve a problem. The researchers operationalized persistence as the "number of times students tried again divided by the total number of incorrect solutions" (p. 310). Students who answered all computations or all application questions correctly were excluded from the persistence data analysis.

The main research procedure was assessing students individually in two separate sessions by one of the researchers. There was a three-month interval between the two testing sessions. In each session, individual students were instructed on testing procedures and then pretested. The researcher explained to the children that she was interested in how confident students felt before, during, and after they solved

mathematics problems. When students understood the instructions and tried a few examples with the researcher, Part One of the OMQ was administered before students started the actual problem solving. While the students worked on each math problem, they were also asked to complete the CDQ, located on the left side of each math problem worksheet. With each incorrect solution, the researcher asked each student whether he or she would like to solve the problem again or go on to the next question. If the student decided to try the problem again, he or she received a new worksheet. No feedback or help was given until the end of the session. At the end of problem solving, students were given Part Two of the OMQ. The researcher recorded the process of each student during both sessions (about 40 minutes per session).

Using repeated measures analyses on students' task-specific appraisals, learning intentions, task performance, and perceived confidence, Vermeer et al. (2000) found some interesting results for computation/application problems and gender effects. For task-specific appraisals and learning intentions, the researchers found a task effect. Student scores of appraisal and learning intentions were higher for computation problems than for application problems. They also found a gender and task interaction. For application problems, boys indicated a higher competence than girls; however, girls perceived higher personal relevance for the task than boys. There was no gender difference in appraisal and learning intentions for computation problems. As for task-specific attributions with application problems, girls attributed failure more to lack of capability and difficulty of tasks than did boys. Again, no gender differences were found with computation problems for attributions.

Regarding task performance, the authors found that boys' mean scores on application problems were significantly higher than girls' mean scores. No gender difference was found on computation mean scores. Examining individual problems also revealed that boys solved all application problems as well as or better than girls did. When comparing performances across tasks, 41% of the students performed better on application problems than on computation problems, while 55% of the students performed better on computation problems than on application problems. Using chi-square statistics, the researchers found a gender difference in performance across types of tasks. They found that more girls performed better than boys on computation problems, while more boys performed better than girls on application problems. For perceived confidence, the researchers aggregated students' scores across phases of the problem-solving process since no variations were found across these phases. Students showed higher perceived confidence for four of the computation problems (e.g., problems 1, 2, 5, and 6) than their paired application problems. A reverse pattern was found for the other two problem pairs (e.g., problems 3 and 4): students' confidence levels were higher for the application than the computation problems. The results revealed that perceived confidence was gender- and task-dependent. Post-hoc t-tests showed a gender difference in perceived confidence with application problems; boys' perceived confidence was significantly higher than that of girls on two of the application problems (e.g., problems 1 and 2). As for persistence after failure experience, a gender difference was again found for application problems. Even though girls scored lower on application problems, they showed higher persistence while solving application problems.

Vermeer et al. (2000) investigated relationships among behavior-related variables such as task performance, perceived confidence, and persistence. Correlation coefficients showed a strong relationship between task performance and perceived confidence for both application and computation tasks. Interestingly, the correlations were weak for persistence and perceived confidence on two types of tasks. The results implied that perceived confidence was inconsistently related to persistence. However, persistence was positively correlated with task performance, suggesting that high performers were more likely to persist than lower performers.

Many findings of the Vermeer et al. (2000) study had valuable implications for mathematics education and an understanding of gender differences concerning students' motivational orientations. This study showed that students had higher task appraisals and learning intentions for computation rather than application problems. Further, gender differences in task appraisal and learning intentions were especially evident in application problems. Boys performed better than girls on application problems and reported having higher confidence during problem solving. Contrary to one of the hypotheses concerning persistence, even though girls showed a lower performance on application problems, they did show a higher level of persistence. Of course, the persistence variable measured in this study did not examine the quality of the effort exerted and that should be of greater importance when assessing persistence in future studies.

Although the study provided many interesting findings and valuable implications, there were several important limitations. First, this study used only 12 problems, so that increasing the number of problems may strengthen some of the results. Second, the

persistence measure was questionable because the task was solvable and thus subject to ceiling effects for skillful learners.

Thus far, the present discussion focused on the well-researched issue of gender and mathematics (Eccles, 1994) and showed, quantitatively, gender differences on mathematics attitude, affect, and performance (Hyde, Fennema, & Lamon, 1990; Hyde et al., 1990). The discussion also centered on gender differences in middle school students' math problem-solving processes and motivational beliefs (Vermeer, Boekaerts, & Seegers, 2000). Examples of interesting findings of this study included: (1) gender differences on mean scores of applied problems (e.g., boys scored higher than girls) but no gender differences on computational problems; (2) girls showed lower confidence than boys did on applied problems; and (3) girls were more likely to attribute negative results to lack of capability and task difficulty. The next section examines the motivational variable of effort and its relationship to self-efficacy and students learning.

Research Studies on Effort and Self-Efficacy

In this section, the discussion focuses on the positive relationship between effort and self-efficacy as well as student performance. Pintrich et al. (1993) developed a self-report instrument, Motivated Strategies for Learning Questionnaire (MSLQ), to assess college students' motivational orientations and learning strategies for a college course. The MSLQ was then tested with 380 undergraduate students at a Midwestern college for its reliability and predictive validity. This instrument contained two main sections: motivational scale and learning strategies scale. Even though the MSLQ is a

comprehensive instrument with 15 subscales, the present discussion here only focuses on (1) the self-efficacy for learning and performance, and (2) effort regulation. The self-efficacy for learning and performance consisted of eight items with questions such as "I believe I will receive an excellent grade in this class," "I'm certain I can understand the most difficult material presented in the readings for this course," "I am confident I can understand the basic concepts taught in this course," and so on. The effort scale consisted of four items with questions such as "I often feel so lazy or bored when I study for this class that I quit before I finish what I planned to do," "I work hard to do well in this class even if I don't like what we are doing," "When course work is difficult, I give up or only study the easy parts," and " Even when course materials are dull and uninteresting, I manage to keep working until I finish" (Pintrich et al., 1991, pp. 13 and 27).

The researchers found the internal consistency of these two subscales to be α = .93 for self-efficacy for learning and performance, and α = .69 for effort regulation. The two scales were correlated with each other (r = .44); most important, each scale was positively correlated with students' course grade (self-efficacy, r = .41; effort, r = .32). It is important to show that a validated instrument assessing students' motivation and learning strategies has positive correlations and also predicts students' performance.

Using the self-efficacy scale from the Motivated Strategies for Learning Questionnaire (MSLQ), Bong (1997a) studied the congruence of measurement on relations between academic self-efficacy, effort, and achievement. The researcher posited that students are assessed in various school subjects with diverse tasks, but are

evaluated based on aggregated (global) performance such as course grades. Thus, the academic self-efficacy measure should have equivalent scope and generality to maximize its predictiveness (Bong, 1997a). She assessed academic self-efficacy in two ways: (1) confidence rating of a sample of problems typical of a school subject (problem-referenced), and (2) the self-efficacy scale of the Motivated Learning Strategies Questionnaire (MSLQ) that assesses students' overall academic confidence in a given course (course-referenced).

With 588 high school students, Bong (1997a) asked students to fill out the self-efficacy scale of the MSLQ, which was slightly modified for high school samples. In addition, students were asked to rate their self-efficacy judgment (i.e., confidence for solving each question correctly) on seven representative questions for six school subjects such as English, Spanish, U. S. history, algebra, geometry, and chemistry. Finally, students reported their most recent grades and filled out a usual effort expenditure scale in each of the six school subjects. The effort expenditure scale consisted of three items per school subject; for example, (1) "How hard do you usually work when studying English?" (2) " How hard do you concentrate when studying English?" and (3) "How much effort do you invest when studying English?" (Bong, 1997a).

The researcher found that problem-referenced and course-referenced academic self-efficacy measures were significantly correlated, ranging from $\underline{r} = .40$ (U. S. history) to $\underline{r} = .72$ (Spanish). Problem-referenced academic self-efficacy was positively and significantly correlated with usual effort expenditure in all six school subjects with correlations ranging from $\underline{r} = .11$ to $\underline{r} = .47$. Further, problem-referenced academic self-

efficacy was positively and significantly correlated with each school subject grade (except U. S. history), with correlations ranging from $\underline{r}$ = .08 to $\underline{r}$ = .60. The course-referenced self-efficacy scale (MSLQ) was also found to correlate positively and significantly with effort expenditure (ranging from $\underline{r}$ = .26 to $\underline{r}$ = .59) and with course grades (ranging from $\underline{r}$ = .26 to $\underline{r}$ = .65). Even though both problem- and course-reference self-efficacy scales correlated positively and significantly with effort expenditure and course grades, the correlations produced from course-referenced self-efficacy had larger magnitudes than problem-referenced self-efficacy. Bong (1997a) found that both self-efficacy scales correlated positively and significantly with effort and course grades. In addition, she advised that the specificity of self-judgment and performance measures should be congruent in order to improve predictiveness of the construct.

These studies by Pintrich et al. (1991, 1993) and Bong (1997a) clearly show that academic self-efficacy positively influences students' effort regulation or effort expenditure and performance such as course grades. To narrow and integrate components of the discussion thus far, the next section examines the mediating role and predictability of mathematics self-efficacy on various outcome measures as well as calibration of self-efficacy in relation to gender.

Research Studies on Mathematics Self-Efficacy

The discussion of this section focuses on studies of self-efficacy in relation to mathematics and accuracy or calibration of mathematics self-efficacy. Hackett and Betz (1989) studied the relationship between mathematical performance and math self-

efficacy, attitudes toward math, and choice of majors by gender. Specifically, they wanted to examine the correspondence between self-efficacy and performance. One of the objectives was to evaluate self-efficacy and performance relationship at a task-specific level. Another objective of the study was to investigate possible differences between males and females in the math self-efficacy and performance correspondence. The last objective was to examine the relationships among variables such as attitudes toward math, math self-efficacy, and math performance, and how these variables predict students' educational choices such as academic major.

Participants in the study totaled 262 (153 women, 109 men) undergraduate students. Students received five instruments during one experimental session after which their ACT scores were obtained from university records. The first instrument was a background and career plans questionnaire used to gather demographic information as well as career plan information. Specifically, three variables (gender, number of years of math taken in high school, and math-relatedness of students' college majors) were important to this study. The second instrument was the Math Self-Efficacy Scale (MSES) with 52 items and three subcomponents: (1) the Math Task Scale (18 items) consisting of items on everyday math tasks such as balancing a checkbook; (2) the Math Courses Scale (16 items) consisting of the number of math courses offered in the college; and (3) the Math Problem Scale (18 items) consisting of arithmetic, algebra, and geometry problems. For the Math Task component, students were asked to rate their confidence level for successfully solving each everyday task. For the Math Course component, students were asked to rate their confidence in completing each course and

their expected grades. For the Math Problem component, students were asked to read each math problem and make a confidence judgment for correctly solving it. The third instrument was a mathematics performance scale that consisted of 18 math questions corresponding to the 18-item Math Problem Scale of the MSES. These 18 questions to be solved were similar, but not identical, to the questions for which students made efficacy judgments. ACT achievement scores (mathematics only) were also obtained as an additional index of math performance. The fourth instrument was the Fennema-Sherman Mathematics Attitude Scale that consisted of 50 items assessing math anxiety, confidence in learning math, perceptions of the usefulness of math, perceptions of math as a male domain, and effectance motivation in math. The fifth instrument was the Bem Sex-Role Inventory (BSRI), which contains a Masculinity scale and a Femininity scale and assesses personality characteristics which are socially desirable for men and women.

Hackett and Betz's (1989) study showed a moderate and positive correlation (r = .44) between self-efficacy and math performance, thereby supporting one of their objectives. To microanalyze the relation between math self-efficacy and performance, the researchers computed deviation scores (D scores) for all 18 items and for each participant. A mean D score was then calculated for each student by averaging the D scores of all items. The D score is a standardized score ranging from -2.00 (underconfidence in performance) to +2.00 (overconfidence in performance). A D score of zero indicates congruence between self-efficacy and performance. The researchers categorized the D scores into five ranges: overconfident (D score > .8); somewhat overconfident (.4 < D score $\leq$ .8); congruent (.4 $\geq$ D score $\geq$ -.4); somewhat

underconfident (-.4 > D score ≥ -.8); and underconfident (D score < -.8). Overall, 35% of students were in the congruent range; however, 48% were in the overconfident range and only 18% were in the underconfident range. When the D scores were broken down into males and females, 43% of women were found to be overconfident and 18% underconfident, while 54% of men were overconfident and 16% underconfident. Even though more men appeared to be overconfident compared to women, a chi-square test showed no statistical difference between the gender and confidence categories.

Hackett and Betz (1989) further investigated math self-efficacy in relation to performance and attitudes toward mathematics. The results indicated that students who scored higher on math self-efficacy, math performance, and math achievement (ACT) tended to report a lower level of anxiety and a higher level of confidence, and saw math as useful, compared with students with lower scores. The researchers also used stepwise multiple regression to test variables that predicted students' college majors. They found that only math self-efficacy, gender, and years of high school math taken were significant predictors of science-/math-related college majors ($\underline{R}^2 = .34$). It seems that male students who reported a higher self-efficacy and took more years of math in high school were more likely to select a science- or math-related major. To further understand the contribution of math self-efficacy in predicting declared college majors, a hierarchical regression analysis was performed since self-efficacy was correlated with math performance variables. The results showed that self-efficacy contributed significantly and uniquely to predicting declared college majors.

In sum, the study by Hackett and Betz (1989) demonstrated the importance of math self-efficacy in predicting the likelihood of declaring a science-/math-related major in college. A most interesting finding of this study was that the majority of students overestimated their capability for solving math problems. Using the Deviation (D) scores, the researchers demonstrated the students' lack of congruence in judging their confidence and actual performance. It would have been interesting if the researchers had included D scores to predict students' declared science-/math-related majors. As such, the researchers were able to establish the important relationship between math self-efficacy and performance. Moreover, they were able to incorporate math self-efficacy and predict students' declared majors beyond math performance and achievement. This study focused on male and female undergraduate students' choices of math- and science-related majors; the next study examines high school students' math- and science-related career interests.

O'Brien, Martinez-Pons, and Kopala (1999) investigated mathematics self-efficacy, ethnic identity, and gender in relation to mathematics and science career interests. The researchers were specifically interested in examining adolescent girls and minority group members for their mathematics and science career interests. They posed three hypotheses: (1) gender and ethnic identity will positively influence students' interests in mathematics- and science-related careers; (2) gender and ethnic identity will influence students' career interests when mediated through mathematics self-efficacy; and (3) mathematics self-efficacy will be influenced by students' past academic performance in the areas of mathematics and science.

The study consisted of 415 (221 male, 194 female) eleventh graders from two high schools. Of these participants, 165 were White, non-Hispanic; 124 were Hispanic; 95 were Black; and 31 were Asian. The researchers used three measures for the study. The first measure was the Mathematics Self-Efficacy Scale (MSES), developed by Betz and Hackett (1983), which assessed: (1) the Math Task Scale (18 items) consisting of items on everyday math tasks; (2) the Math Courses Scale (16 items) consisting of the number of math courses offered in high school; and (3) the Math Problem Scale (18 items) consisting of arithmetic, algebra, and geometry problems. For Math Task, students were asked to rate their confidence level for successfully solving each everyday task. For Math Course, students were asked to rate their confidence in completing each course with a grade of B or better. For Math Problem, students were asked to read each math problem and make a confidence judgment. The second measure was the Multigroup Ethnic Identity Measure (MEIM), which assessed student ethnic attitudes and sense of belonging, ethnic identity achievement, and ethnic behaviors or practices. The third measure was adapted from the Jackson Vocation Interest Survey (JVIS) to assess students' career interests in engineering and science. The researchers also obtained SES information and PSAT scores from participants.

Results of the study showed a number of significant correlations: (1) career interest with self-efficacy, PSAT, and gender; (2) self-efficacy with PSAT and ethnic identity; and (3) PSAT with SES. The researchers conducted further analyses using a path model. The path model explained 24% of the variance in interest in pursuing science and engineering careers, on which self-efficacy and gender had direct effects.

This model accounted for 15% of the variance in mathematics self-efficacy, as contributed by PSAT and ethnic identity. The model also explained 2% of the variance in PSAT, contributed entirely by SES. The path analysis showed a significant association between self-efficacy with a career interest in science and engineering, self-efficacy and prior academic achievement (PSAT), and PSAT-influenced career interest as mediated by self-efficacy.

In sum, this study demonstrated that mathematics self-efficacy was a significant predictor of mathematics and science career interests among adolescents. Self-efficacy was further predicted by past academic achievement and ethnic identity. As the researchers hypothesized, gender directly predicted that students' career interests in science and mathematics favored boys. Further, this study demonstrated that self-efficacy was influenced by ethnic identity and past academic performance. The findings also suggested that the mediating roles of past academic performance and mathematics self-efficacy are important intervening factors that increase the interest of girls and minority students in the fields of mathematics and science. However, O'Brien, Martinez-Pons, and Kopala's (1999) study has limitations. Even though ethnic identity influenced career interest as mediated by self-efficacy, it was not clear which ethnic groups were more likely to be interested in mathematics- and science-related careers. Similarly, it was not clear from the path analysis whether a gender x ethnic identity interaction influenced career interests.

Thus far, the present discussion has centered on students' math- and science-related career interests and choice of majors, and how these outcomes were influenced by

perceived self-efficacy, among other variables. The next study tested the mediating role of self-efficacy and compared it to other motivational variables, such as self-concept, by using path analysis.

Pajares and Miller (1994) set out to test Bandura's hypotheses regarding the predictive and mediational role of self-efficacy in mathematics through path analysis. Specifically, the researchers wanted to determine whether students' self-efficacy judgments had a stronger direct effect on their problem-solving performance than did other variables (such as self-concept, anxiety, perceived usefulness, and prior experience) as related to mathematics and gender. Further, they examined whether self-efficacy mediated the effect of gender and prior experience on these variables and problem-solving performance.

In this study, 350 undergraduate students (229 female, 121 male) from a public university in the South participated. Students filled out a number of scales during individual classes. First was a Mathematics Confidence Scale (MCS) which measured math self-efficacy. Second was a 20-item instrument that measured perceived usefulness of mathematics in domains such as social activities, employment, education, family life, and citizenship. Third, the Mathematics Anxiety Scale, a 10-item scale adapted from Betz (1978), was used to measure math anxiety. Fourth, the Self-Description Questionnaire, adapted from Marsh (1992) and consisting of academic and course-specific self-concepts, was used to measure math self-concept. Fifth, a self-report of the maximum level of math courses attained in high school was used to measure students' prior experience with math. Sixth, a self-report of semester credits earned in

mathematics was used to assess students' college math experience. Seventh, the Mathematics Problems Performance Scale, an 18-item multiple-choice test, was used to measure students' math performance.

Using path analysis, Pajares and Miller (1994) showed that self-efficacy had significant direct effects on all other variables. Further, both direct and total effects of self-efficacy were significantly stronger than other variables in the path model. Participants' judgments of their capability to solve math problems were more predictive of that problem-solving ability than were other factors in this analysis. As for gender, its effect on self-concept and performance was largely indirect and mediated by the self-efficacy variable. Male students reported higher math self-efficacy than did female students. Men also scored higher on performance than did women, even though no differences were found on prior math experience. Female students also reported higher anxiety than their male counterparts. Path analysis showed that differences in math self-efficacy contributed to differences in math performance.

The researchers found that, overall, students overestimated their math performance competence. Overestimation was defined as rating a question 4 or 5 (much confidence or complete confidence) on the self-efficacy scale but incorrectly solving it. Underestimation was defined as rating a question 1 or 2 (no confidence at all or very little confidence) but correctly solving it. Pajares and Miller (1994) found that 57% of students overestimated their performance (female, 57%; male, 58%), while 20% underestimated their performance (female, 21%; male, 17%). However, the researchers

found no significant difference between males and females on either overestimation or underestimation of math competence.

The results of this study supported the authors' hypotheses regarding the perceived and mediational roles of math self-efficacy. Math self-efficacy was more predictive of math performance than other motivation and math-related variables that was studied. Math self-efficacy was found to be an important variable that mediated the effect of gender and prior math experience on self-concept, perceived usefulness of math, and math performance. Pajares and Miller's (1994) study demonstrated the importance of self-efficacy for influencing students' math performance above and beyond numerous motivation and math-related variables.

Thus far, this review has discussed the importance of the self-efficacy construct and its mediating and predictability roles in mathematics performance, choices of math- and science-related majors and career interests, all in relation to gender. However, as the present study attempted to investigate math self-efficacy and calibration of self-efficacy in relation to gender, the focus of the review is now narrowed to specific studies that examined these particular topics.

Ewers and Wood (1993) conducted one of the first studies to examine the accuracy of self-efficacy judgments. The purpose of this study was to investigate gender and ability differences between gifted and average-ability fifth graders in mathematics self-efficacy and prediction accuracy. The researchers focused on two self-efficacy dimensions: level and strength. An equal number of students (n = 19) comprised each group: average female, average male, gifted female, and gifted male. The criterion for

average-ability students was based on a general aptitude score (Cognitive Abilities Test) which fell between 89 and 111. The gifted students were selected from the gifted program.

Students were asked to make self-efficacy judgments on 20 math problems that consisted of 5 easy, 10 moderate, and 5 difficult problems. The students were asked to examine each math task and respond to the corresponding certainty question, "How sure are you that you will be able to correctly solve this problem?" The certainty scale was a 5-point scale, with responses ranging from not at all sure (1) to totally sure (5). After students made all of their predictions, they were instructed to solve the math problems. Self-efficacy level was measured by counting the number of problems out of 20 (the total number of problems) that had certainty ratings above 3; they calculated self-efficacy strength by summing the ratings and dividing by 20.

Using a 2 x 2 (gender by ability) ANOVA, the researchers found a statistical difference between the gifted and average groups on math performance, but no statistical difference between gender groups. The gifted group indicated a significantly higher level of math self-efficacy than the average group. However, no significant gender differences were found in math self-efficacy levels. Another 2 x 2 (gender by ability) ANOVA analyzed self-efficacy strength. The researchers found gender and ability main effects: Gifted students showed higher self-efficacy strength than average students. Further, regardless of ability level, males showed significantly higher self-efficacy strength than females, but there were no significant interactions.

Students' prediction accuracy was investigated by calculating the number of negative hits for high self-efficacy predictions (ratings of 4 or 5 for corresponding problems solved incorrectly). A 2 x 2 (gender by ability) ANOVA showed ability main effect and a marginal gender effect. The gifted group made fewer overestimations of their ability to solve math tasks than the average group, but male students showed more overestimations than female students. Another measure of prediction accuracy involved the number of positive hits for low self-efficacy predictions (ratings of 1 or 2 for corresponding problems missed). A 2 x 2 (gender by ability) ANOVA showed an ability main effect: The average group made fewer underestimations than the gifted group. In other words, gifted students correctly answered more problems even though they initially judged that they would not be able to solve them.

In sum, Ewers and Wood's (1993) study demonstrated ability and gender differences in math self-efficacy among fifth graders, but no gender differences in math performance. As for prediction accuracy, gifted students made fewer overestimations than regular students, and girls made fewer overestimations than boys. However, average students were more accurate than gifted students regarding low self-efficacy predictions. This study was an early attempt to address calibration of self-efficacy and performance; the researchers attempted to measure calibration as prediction accuracy and operationalize it as negative and positive hits. This study shed light on ability and gender differences on math self-efficacy judgments and calibration.

Another study by Pajares and Kranzler (1995) examined the influence of math self-efficacy and general mental ability on the math problem-solving ability of high

school students. The purpose of their study was to examine the predictive and mediation role of self-efficacy, while controlling general mental ability, on math problem solving in the high school environment. They hypothesized that (1) self-efficacy mediates the effect of gender, math level, and ability on math anxiety and math problem solving, and (2) self-efficacy independently contributes to predicting math problem-solving outcomes.

To test these hypotheses, the researchers first asked 329 high school students (grades 9-12) to complete the Raven's Advanced Progressive Matrices (APM) general mental ability test. Second, students were asked to make math self-efficacy judgments on 18 mathematics items consisting of arithmetic, algebra, and geometry problems. Third, students were asked to answer 10 math anxiety items consisting of five positively and five negatively worded items. Last, students were asked to solve the 18 math problems from which their math self-efficacy scores were obtained. In addition to these measures, the researchers included students' math levels as another variable in the study. Math levels were operationally defined as students' highest math course (1 – applied math; 7 – calculus) completed by the time of their participation in the study. The researchers attempted to study calibration by examining students' overconfidence and underconfidence. Overconfidence was defined as students' self-efficacy rating (4, 5, or 6 out of a possible 6) on an item with an incorrect solution; underconfidence was defined as a rating of 1, 2 or 3 with a correct solution.

Using path analysis, Pajares and Kranzler (1995) examined the direct and indirect effects among variables in order to answer their hypotheses. One of the important findings from path analysis was that self-efficacy and ability have a strong direct effect

on math problem solving. Furthermore, self-efficacy has a strong effect on math anxiety, demonstrating that math anxiety was a by-product of low math efficacy judgment. Path analysis showed a moderate effect of math level on self-efficacy, suggesting that students' experience in math courses are important sources of self-efficacy information. The path model showed that gender had no significant direct effect on self-efficacy; however, inclusion of this path improved the model fit. Thus, gender has a direct and significant effect on math anxiety.

The researchers also found that most high school students were overconfident in their capability to solve math problems. According to their study, 86% of the high school students overestimated and 9% underestimated their math performance. Only 1% ($\underline{n} = 4$) of the sample accurately predicted their responses on all 18 items. Interestingly, they found that students who were overconfident erred more often (an average of 6.2 problems) than students who were underconfident (an average of 3.5 problems).

Because ethnicity was not a variable used in the path model, the researchers performed a planned pairwise t-test comparison of ethnicity on variables using the path model. They found a few differences between African-American students and non-Hispanic White students. First, African-American students had significantly lower math performance than their counterparts. Second, African-American students also rated their self-efficacy significantly lower than White students. Interestingly, more African-American students were overconfident about their math capability despite their lower math performance. Significant differences were found between African-American (6.8 problems) and White (5.2 problems) students on the number of overconfidence items.

In sum, Pajares and Kranzler (1995) demonstrated that the math self-efficacy of high school students played a predictive and mediational role in math problem solving. Using path analysis, the researchers also demonstrated the direct effect of math self-efficacy on math anxiety and performance when general mental ability is controlled. In this sample, most of the participating high school students, regardless of gender, were overconfident of their math capability. However, students who attained higher math levels were more accurate in judging their math capabilities. Although Bandura (1986) stated that overestimates of one's capability can sustain personal effort and persistence, the question remains: how much overconfidence is useful and how much becomes detrimental? Other questions are: why are students overconfident and what are the sources of such overconfidence?

A later study by Pajares and Miller (1997) examined the forms of assessment that influence students' math self-efficacy judgments and the relationships between math self-efficacy and performance. Specifically, an objective of the study was to determine whether multiple-choice versus fill-in-the-blank forms of assessment influenced students' math self-efficacy judgments. A second objective of the study was to determine the relationship between math self-efficacy judgments and performance (calibration) as related to the two forms of assessment. In addition, the researchers were interested in how gender may influence math self-efficacy, because this link had not been thoroughly studied previously. Three rationales for examining assessment formats included understanding: (1) the effects of the format of math problems on confidence judgments;

(2) the implications of these formats for learning and high-stake assessments, such as statewide tests; and (3) the effects of assessment formats on calibration.

The study included 327 eighth graders (178 girls, 149 boys), 199 of whom were in algebra and 128 in pre-algebra. Students were divided into four groups: Groups 1 and 3 made self-efficacy judgments on open-ended math questions, while Groups 2 and 4 made self-efficacy judgments on similar math questions with multiple-choice answers. After students made their self-efficacy judgments on the questions, they were given the math problems in two formats: open-ended and multiple-choice. Groups 1 and 2 received multiple-choice while Groups 3 and 4 received the same instrument with an open-ended format. The four groups were labeled as follows: OE Eff/MC Perf (Group1); MC Eff/MC Perf (Group2); OE Eff/OE Perf (Group3); and MC Eff/OE Perf (Group4).

An instrument used in this study was a single-rating self-efficacy scale that accompanied each of the 30 math questions. This 6-point Likert-type rating scale (1 = no confidence at all; 6 = complete confidence) measured students' confidence level in correctly solving each math problem. The math performance instrument consisted of the same 30 problems that were used to make self-efficacy judgments. The OE Eff/OE Perf and MC Eff/OE Perf groups of students received the math test in a fill-in-the-blank version, while the OE Eff/MC Perf and MC Eff/MC Perf groups received a multiple-choice version of the test. The multiple-choice version consisted of the same questions with five possible answers for each question; students were asked to circle one correct answer per question. To assess students' calibration, Pajares and Miller (1997) used three measures. The first measure was mean bias that revealed the direction of the error

judgments; this was calculated by subtracting actual performance from self-efficacy judgment. The second measure was mean accuracy that revealed the magnitudes of the error judgments; this was calculated by subtracting the absolute value of each bias score from the absolute value of the highest range score. The third measure was item accuracy that reflected the number of math questions for which the self-efficacy judgments and performance scores were in accord. For example, an item accuracy score was the number of questions for which students rated a confidence level of 4, 5 or 6 and then correctly solved, plus the number of questions with a confidence rating of 1, 2 or 3 and then incorrectly solved.

Using a MANOVA, Pajares and Miller (1997) showed a significant between-group effect for dependent variables such as math performance, self-efficacy, mean bias, mean accuracy, and item accuracy. To test for group effects on all dependent variables, the researchers used an ANOVA and found that students did not differ in self-efficacy judgments, but differed in performance, mean bias, mean accuracy, and item accuracy. First, students did not differ in their confidence level, regardless of the problem format presented to them. However, the two groups of students who took the multiple-choice performance test significantly outperformed those who took the open-ended question test. Thus, the groups that took the open-ended tests displayed larger mean bias, revealing greater overconfidence.

Using multiple regression, Pajares and Miller (1997) analyzed predictions of math performance. They found that the form of self-efficacy assessment did not predict math performance. In other words, students who rated their self-efficacy with an open-ended

format (OE Eff) did not differ in math performance, compared with students who rated their self-efficacy (MC Eff) with a multiple-choice format. As researchers expected, there was a main effect for performance format. Students who took the multiple-choice format (MC Perf) test scored significantly higher than those who took the open-ended format (OE Perf) test. Further, math self-efficacy and math level predicted students' math performance. Students enrolled in the algebra class who demonstrated a higher confidence level correctly solved more problems than students with a lower confidence level who were enrolled in the pre-algebra class. Regarding the students' math performance predictions, the researchers found a gender by self-efficacy level interaction. Interestingly, this interaction indicated that girls with low self-efficacy outperformed boys with low efficacy; however, boys with high self-efficacy outperformed girls with high self-efficacy.

As for calibration, it was found that students who were enrolled in algebra and pre-algebra differed significantly not only in performance, but also in mean bias, mean accuracy, and item accuracy. Interestingly, students in algebra or pre-algebra did not differ in their self-efficacy levels. They also found that, regardless of self-efficacy formats, students who were later tested with an open-ended question format were more poorly calibrated than students who tested with a multiple-choice format. As for gender effect, the researchers found no significant gender differences in performance and calibration.

Pajares and Miller's (1997) study attempted to determine whether a traditional assessment format (multiple-choice), compared with another format (open-ended),

influenced students' self-efficacy judgments. The results showed that student self-efficacy was not affected by different assessment formats. However, the calibration of some groups of students (MC Eff/OE Perf, OE Eff/MC Perf) was poorer than others because the former groups were not informed of changing formats for subsequent tests. As a result of these format changes, the students' level of calibration could have been adversely affected. The researchers concluded that an open-ended assessment format could provide a more accurate measure of students' math capability than a multiple-choice format. Math performance with an open-ended format would not be inflated by guessing, thus worsening calibration due to students' overconfidence.

Pajares and Miller (1997) found no gender differences on math self-efficacy, performance, and calibration measures. However, with higher self-efficacy scores, boys were better predictors of their math performance than their female counterparts. By contrast, reverse relations occurred with lower scores of self-efficacy; boys were poorer predictors of their math performance than were girls.

Pajares and Graham (1999) studied the influences and changes of various motivation constructs on the math performance of middle school students. Their research objectives were to: (1) determine whether math self-efficacy independently contributed to the prediction of math performance when other motivation variables and previous achievements were controlled; (2) discover when math self-efficacy began to change during the first year of middle school; and (3) examine whether gender and regular/gifted education placement varied on a number of motivation constructs. The motivation constructs examined in this study were: math self-efficacy, math anxiety, math self-

concept, self-efficacy for self-regulated learning, value judgment of math, and engagement.

The researchers studied 273 sixth graders (123 girls, 150 boys), including 188 regular and 85 gifted students. Participants were given the measures of motivation and math performance at the beginning (October) and toward the end (April) of the academic year. One measure of math self-efficacy was a task-specific assessment of students' perceptions of their competence for solving each math problem correctly. The 8-point scale ranged from 1 (not confident at all) to 8 (completely confident). To measure students' math anxiety, the researchers adapted 8 out of 10 items from Betz's Mathematics Anxiety Scale (MAS), on which high scores indicate high anxiety. The math self-concept measure contained six items adapted from the Academic Self-Description Questionnaire II (ASDQII) (Marsh, 1990, 1992), which measures students' global attitudes and perceptions toward mathematics (e.g., "I have always done well in mathematics"). They also assessed students' self-regulated learning and strategies used with a sub-scale of Bandura's Children's Multidimensional Self-Efficacy Scale (Zimmerman, Bandura, & Martinez-Pons, 1992). The researchers used 7 of the 11 items that focused on assessing students' capability judgments about finishing homework on time, planning schoolwork, and studying when faced with distractions. The researchers also measured students' value judgments about mathematics such as perceived importance, interest, and enjoyment. To assess students' engagement, Pajares and Graham used three items to measure their effort and persistence in math.

In addition to motivation-related variables, outcome measures of students' mathematics performance, previous academic achievements, and calibrations were included. The math outcome measures were end-of-unit exams similar to those on which students made their math self-efficacy judgments. Previous achievements included students' percentile scores on the Iowa Test of Basic Skills, and students' grade point averages in mathematics for semesters during fifth and sixth grades. To measure calibration of self-efficacy, Pajares and Graham used two measures: mean bias and mean accuracy, as recommended by Schraw (1995). Mean bias, indicating the direction of the error judgments, is calculated by subtracting the actual performance from the self-efficacy judgment. Mean accuracy, indicating the magnitudes of error judgments, is calculated by subtracting the absolute value of each bias score from the absolute value of the highest range score.

Pajares and Graham (1999) conducted multiple regression analyses to determine whether self-efficacy made an independent contribution when other variables were controlled. Separate analyses were done for data collected at the beginning and end of the academic year. Simple t-tests were also calculated to determine differences on measures taken in October and April. Finally, multivariate analyses were conducted to test whether any observed differences in math performance, motivation-related variables, and calibration were due to gender, regular/gifted placement, and their interactions.

The results showed that the April math performance test was more difficult than the October test. In accordance with math performance, math self-efficacy scores were lower in April than in October. As predicted, gifted students outperformed regular

students on both math tests, but boys and girls demonstrated no gender differences in math performance tests. Pajares and Graham (1999) also found that the independent variables accounted for 56% (October) and 53% (April) of the variance in math performance. To fulfill the first objective of the study, the researchers compared differences between the full model (with self-efficacy as a predictor of performance) and the reduced model (removing self-efficacy as a predictor of performance). The results showed that self-efficacy made an independent, but modestly significant, contribution to predict math performance in both October and April (Pajares & Graham, 1999).

To achieve the second objective, discovering the changes in math-related constructs at the beginning and end of the academic year, simple t-tests with adjustment for multiple comparison were conducted. The researchers found significant differences between the two semesters on math performance and self-efficacy scores, but not on math anxiety, self-concept or self-efficacy for self-regulation. Interestingly, the researchers found that students' value of math decreased, and they reported exerting lower effort and persistence at the end rather than the beginning of the academic year. Further, students were more biased and overconfident at the end of the school year; their self-efficacy beliefs were less in accord with their performance scores.

As for gender and giftedness differences on math and motivation-related variables, the researchers found no gender differences in math performance (for both fall and spring). There was also no gender difference in any of the motivational constructs. However, they found that gifted students had higher math performance scores and were less biased toward overconfidence. Gifted students were better calibrated than regular

students and also reported higher math self-efficacy and math self-concepts for both October and April administrations.

Although the main objectives of Pajares and Graham's (1999) study did not include calibration (the accuracy of students' self-perceptions and their actual performance), their findings that students' calibration measures were highly correlated with math performance were intriguing. Overall, students were overconfident in their self-efficacy judgments, but gifted students had less overconfidence than regular students. The mean accuracy measure of calibration correlated (fall, $\underline{r}$ = .88; spring, $\underline{r}$ = .81) with math performance, exceeding the uncalibrated measure of self-efficacy (fall, $\underline{r}$ = .57; spring, $\underline{r}$ = .59). This suggests that calibration may improve the prediction validity of self-efficacy measures. However, Pajares and Graham's (1999) approach for assessing calibration had limitations, namely that students' math performance scores were also part of the two measures of calibration. Thus, the question of the value of self-efficacy calibration remains unanswered.

The following study used different methods of measuring confidence judgments of undergraduate students on test questions. Lundeberg, Fox, and Puncochar (1994) examined the confidence of undergraduate students after they answered specific test questions. The purpose of their study was to examine the potential gender differences among undergraduates in making confidence judgments on course materials. The researchers were interested in answering two key questions: (1) "Are men more confident than women that their answers to exam questions are correct?" and (2) "Are men better calibrated in confidence than women; that is, do higher confidence ratings

indicate appropriate accuracy and lower rating inaccuracy?" (p. 115). Furthermore, the researchers posed follow-up questions which: (1) tapped into possible gender differences in specific content items, (2) compared top-half and bottom-half students' confidence, and (3) compared graduate and undergraduate students' confidence.

Participants (181 women, 70 men) were drawn from three different psychology courses. Two groups of students came from psychology laboratory methods sequence courses (Lab 1 and Lab 2); the third group came from a learning and memory course that consisted of graduate and upper-division undergraduate students (Memory Group). All students in the three groups took pretest and final exams. Pretests consisted of quizzes given during classes. After answering each question, students were given a corresponding rating scale on which to indicate their confidence in answering the questions correctly. The Lab Groups rated their confidence on a 5-point scale ranging from pure guess (1) to very certain (5). The confidence estimates of the Memory Group ranged from 50% to 100% for true-false items and 25% to 100% for multiple-choice items. For final exams, Lab 1 students had to answer 27 multiple-choice questions accompanied by 27 confidence ratings. Lab 2 students had to answer 23 multiple-choice questions along with 23 confidence ratings. Specific domains assessed in Lab 1 and 2 final exams were computational skills, experimental design, descriptive statistics, and conceptual content (e.g., animal learning, auditory psychophysics, and reaction time). Memory Group students had 23 true-false and 17 multiple-choice questions along with confidence estimates. Lundeberg, Fox, and Puncochar (1994) only presented data collected on the final exams.

For calibration of participants' confidence accuracy, the researchers calculated students' Confidence Accuracy Quotient (CAQ): mean confidence correct minus mean confidence wrong, divided by standard deviation of confidence (across both correct and incorrect). A zero CAQ indicated that an individual had no calibration of confidence. A positive CAQ score indicated higher confidence when correct than when wrong, and a negative CAQ score indicated higher confidence when wrong than when correct.

The results showed that in both Lab 1 and Lab 2, men's confidence levels were higher than women's when answering questions correctly and incorrectly. The Memory Group indicated no gender differences regarding confidence level, but both genders in this group overestimated their likelihood of being correct. The researchers found, in general, that both men and women showed a moderate degree of calibration in their confidence ratings. In other words, participants expressed higher confidence when answers were correct than when answers were wrong. Significant gender differences in CAQ scores were found in Lab 1 and Lab 2, with women obtaining higher scores than men. Women in the two groups indicated higher confidence when answers were correct than when answers were wrong. The Memory Group showed no significant gender differences in CAQ scores.

When examining domain-specific gender differences in both Lab groups, the results revealed significant gender differences in confidence with items assessing math computational skills. Male students displayed higher confidence than female students did when correct and incorrect. Similarly, with descriptive statistics questions, male students in Lab 2 were more confident than their female counterparts when the answers were

wrong. However, men's higher confidence levels were not shown to be consistent across all four content areas. Gender differences in confidence levels were dependent both on contexts and on domain being tested. In the domain of math computation, men were more confident than women, whereas in other domains (psychological science and experimental design), there were no significant gender differences. Calibration in specific contexts was dependent on gender in the two Lab groups. For example, women calibrated 87% of the time for 7 out of 8 instances (2 Labs x 4 domains), and men calibrated 37% of the time for 3 out of 8 instances. In general, the researchers noted that when women were wrong, their confidence ratings were closer to 3 (mixed feelings of confidence and uncertainty), whereas when men were wrong, their confidence ratings were closer to 4 (reasonably confident) (Lundeberg, Fox, & Puncochar, 1994).

Both men and women in the Memory Group showed overconfidence in the accuracy of their answers, but women in this group were slightly but not significantly more confident than men. Since the Memory Group consisted of both graduate and undergraduate students, separate analyses were also conducted. Among graduates in the Memory Group, women were more confident than men about answering correctly and incorrectly. As for the undergraduates in the Memory Group, undergraduate men were more confident when incorrect than when they were correct.

Overall, Lundeberg, Fox, and Puncochar's (1994) study showed that although both women and men were generally overconfident, undergraduate men were especially overconfident when incorrect. Further, the gender differences of confidence and calibration were dependent on the content of the test items. In some content areas such as

math computation and statistics, men were more confident than women; in other domains such as experimental design and psychological sciences, however, no gender difference was found. Nevertheless, the methodology of Lundeberg, Fox, and Puncochar's (1994) study had its limitations. First, students rated their confidence judgments after they had answered the questions; it was a post-hoc study. The process of solving questions before making confidence judgments provides more information than self-efficacy judgments made prior to solving problems. Thus, results may differ if the confidence judgments were made prior to answering the questions. Second, the two Lab groups (Lab 1 and 2) did not answer the same number of questions and the Memory Group assessed their confidence level differently, making interpretations of the results more difficult.

In summary, this section of the literature review examined studies that tested the mathematics self-efficacy construct in relation to a number of cognitive and motivational variables. Hackett and Betz (1989) and O'Brien, Martinez-Pons, and Kopala (1999) investigated mathematics self-efficacy in relation to math- and science-related choices of majors and interests. Both studies showed that math self-efficacy is an important motivational variable that positively predicts college and adolescent students' choices of science- and math-related majors. The discussion then examined the mediating role of math self-efficacy using path analysis (Pajares & Miller, 1994). Compared to self-concept and other math-related variables, self-efficacy was more predictive of math performance. Ewers and Wood's (1993) study was one of the first to examine the accuracy judgments of math self-efficacy. Pajares and his colleagues (Pajares & Graham, 1999; Pajares & Kranzler, 1995; Pajares & Miller, 1997) refined the study of the self-

efficacy accuracy of calibration, using measures such as mean bias and mean accuracy. Finally, Lundeberg, Fox, and Puncochar (1994) studied the confidence of undergraduate male and female students by using measures that differed from Pajares and his colleagues.

The next and last section of this chapter presents a synthesis of the research discussed in the literature review as a foundation and rationale for the present study. Drawing from these research studies, the hypotheses were then asserted and tested in the present study.

Synthesis and Hypotheses

To synthesize research on the calibration of mathematics self-efficacy that was discussed in the literature review, it is crucial to reanalyze the pertinent points of each sub-section. The sub-sections consisted of an overview of self-efficacy theory, the uniqueness of the self-efficacy construct in relation to academic achievement and motivation, measurement issues of calibration, mathematics in relation to gender, research on self-efficacy and effort, and studies on mathematics self-efficacy.

A theoretical foundation of self-efficacy by Bandura (1986, 1997) anchored the issues of calibration of mathematics self-efficacy. Bandura (1986, 1997) introduced the construct of self-efficacy, which refers to one's judgments of his or her capability to perform diverse tasks at designated levels. Self-efficacy affects how an individual makes choices in his or her environment as well as choices of activities, of courses of action to take, of amount of effort to exert, and of length of time engaged on each task. Bandura

(1997) asserted that self-efficacy has three dimensions—level, strength, and generality—it should be assessed in terms of "can do" rather than "will do." Further, self-efficacy should be measured in correspondence to the tasks in question while the tasks themselves should be comprised of various difficulty levels. The theory of self-efficacy has been widely studied in many academic fields including mathematics.

After establishing the theoretical foundation, the focus was on the uniqueness of the self-efficacy construct that was summarized by Zimmerman (2000) and quantified by Multon, Brown, and Lent (1993). Zimmerman (2000) discussed its uniqueness and positive influence on academic learning, achievement, motivation, and self-regulation processes. Research showed that students with higher self-efficacy were more likely to put forth effort when encountering adverse situations, choose more challenging activities, and increase persistence when engaging activities. A high sense of self-efficacy promotes greater motivation to learn; students become more engaged in self-regulated learning and, in turn, produce higher academic performance. The researcher pointed to key research studies that demonstrated the uniqueness of the self-efficacy construct and its positive influence on student motivation and academic learning. To provide further evidence of the validity of the self-efficacy construct in relation to academic achievement and motivation (e.g., persistence), Multon, Brown, and Lent (1993) conducted a meta-analysis to quantify the effects of research studies pertaining to academic self-efficacy. They found that self-efficacy had an overall effect size of .38 and accounted for 14% of the variance in students' academic performance. As for self-efficacy in relation to academic persistence, the researchers found an overall effect size of .34 while self-

efficacy accounted for 12% of academic persistence. It is clear that self-efficacy is an important construct in relation to student academic achievement and possesses a mediating role of academic motivation. Even though self-efficacy, according to Multon et al. (1993), accounted for 14% of the variance in academic achievement and 12% in motivation (e.g., persistence), much variance has not been explained. It may be possible to increase the predictability of self-efficacy to explain academic achievement and academic motivation. Although a higher sense of self-efficacy improves student achievement, it is not clear whether a more accurate sense of self-efficacy contributes above and beyond to self-efficacy levels and strength.

To understand self-efficacy accuracy or calibration, its measurement issues were considered. Schraw (1995) distinguished two ways to measure feeling-of-knowing in terms of agreement accuracy and association. Agreement accuracy was the degree of feeling-of-knowing judgment of an event and its observed outcome. Measuring self-efficacy calibration aligned closely with measuring agreement accuracy. The researcher recommended measuring agreement accuracy in terms of its bias and accuracy. These two measures were based on the Euclidean distance between predicted and observed performance. Thus, bias measures the direction of judgment errors while accuracy measures the magnitude of judgment errors. Using these two measures as part of measuring calibration, Schraw, Potenza, and Nebelsick-Gullet (1993) studied the effects of feedback and external incentives on academic performance and calibration. They found students to be better calibrated and demonstrating lesser overconfidence with the easy test than with the difficult test. Further, students who received external rewards to

improve calibration did better on their performance and were less biased and more accurate than students who received rewards for improving performance.

Before shifting the focus to calibration of mathematics self-efficacy, measurement issues of calibration need to be examined. The discussion now centers on important issues of mathematics and gender. Mathematics in relation to gender has mostly been examined from the perspectives of career options, occupational interest, and educational choices. Eccles (1994) summarized numerous studies on gender in relation to educational and occupational choices. She and her colleagues specifically investigated motivational and social factors that influence students' achievement goals, career interests and choices, course selections, persistence and effort. Eccles' (1994) achievement-related model was developed based on decision-making models, achievement theories, and attribution theories. Eccles concluded that female students were less likely to choose fields such as mathematics, science, and engineering than male students. She also found that females had less confidence in their abilities and placed less value on math-related fields than males did. After examining differences on gender in relation to mathematics, the discussion then centered on quantifying these differences. Hyde, Fennema, and Lamon (1990) conducted a meta-analysis to test gender differences in mathematics performance. They found, overall, that gender difference was small in math performance; however, more noticeable gender differences in math performance were found when considering participants' age, selectivity of sample, and cognitive level of the test. In another meta-analysis, Hyde et al. (1990) compared gender and mathematics attitudes and affect. Overall, they found small effect sizes with one

exception: the stereotyping of math as a male domain. This finding indicated that more males than females believed math to be a male domain. When the researchers examined age and gender differences in math attitudes and affect, they found larger effect sizes for the high school-age group; males had more positive math attitudes as their age increased, compared to females. Overall, effect sizes were small, but some were worth noticing. The authors also found positive effect size (favoring males) for math self-concept and confidence, with a larger effect size for the high school-age group. Again, for math anxiety, males had lower anxiety compared to females. Even though Hyde et al.'s (1990) meta-analysis produced, overall, small effect sizes on gender differences on attitudes and affect, any statistically significant effect sizes were still worth noting.

Recently, Vermeer, Boekaerts, and Seegers (2000) conducted a study to examine middle school students' motivation and gender differences. Specifically, the researchers were interested in male and female student differences in cognitive and affective changes before, during, and after mathematics tasks (computations and applications). They found no gender differences on solving computational tasks, but some gender differences on solving applied problems (boys scored higher than girls). When examining perceived confidence, the researchers found that students had higher confidence with computational tasks than with applied tasks. However, they also found girls to have lower confidence on applied tasks compared to boys; girls were more likely to attribute their failure to lack of capability and to task difficulty. Girls also showed a higher persistence rate than boys on applied tasks. Vermeer et al.'s (2000) study is a recent effort to investigate gender differences on math problem-solving behaviors and motivation. Even though this study

was conducted in the Netherlands, it lends support for further research on gender differences in mathematics problem-solving behaviors and achievement motivation. Next, the discussion focuses on the relationship between self-efficacy and motivational variable, effort, and its contribution to predicting performance.

Pintrich et al. (1991, 1993) developed the Motivated Strategies Learning Questionnaire (MSLQ) to assess students' motivational beliefs and learning strategies. Working with college population, they found that students' self-efficacy was positively and significantly correlated with effort regulation. In addition, effort positively influenced students' course grades. It was clear that effort is an important motivational variable and highly correlated with self-efficacy. Bong (1997a) compared the self-efficacy measure from the MSLQ (which she termed course-referenced self-efficacy) and item specific self-efficacy scale (which she termed problem-referenced self-efficacy) to test the predictiveness of students' effort expenditure and course grades. She found that the course-referenced self-efficacy measure had a higher and significant correlation with students' effort expenditure as well as their course grades. She cautioned researchers that measures of students' self-perceptions such as self-efficacy and performance such as course grades should be congruent with their globalness or specificity. Next, the discussion interweaves the self-efficacy theory, research studies on calibration, and issues on gender and mathematics.

Mathematics self-efficacy has been shown to predict math performance; it is more highly predictive than other common motivation and math-related variables (Pajares & Miller, 1994). In addition to being a unique and important construct in achievement

motivation, a great contribution of self-efficacy was an understanding of gender differences in the occupational and educational fields of mathematics. While math self-efficacy has great predictability for math performance, it also predicts the likelihood of college students choosing majors in the field of mathematics and science (Hackett & Betz, 1989). Furthermore, math self-efficacy is a good predictor of high school students' interests in pursuing math and science careers (O'Brien, Martinez-Pons, & Kopala, 1999).

The focus of self-efficacy studies also shifted to an examination of self-efficacy calibration—the accuracy of students' self-efficacy and their actual performance. Ewers and Wood (1993) examined the accuracy of self-efficacy judgments of fifth graders. They found gifted students to be more accurate about their capability for solving math tasks than average students. They also found that boys made more overestimations of their capability to solve math problems than did girls. However, their measure of accuracy was not optimally sensitive. They counted the number of negative hits for high self-efficacy predictions (rating 4 or 5 for corresponding problems solved correctly) and the number of positive hits for low self-efficacy predictions (rating 1 or 2 for corresponding problems missed). Such measurements of accuracy were not sensitive enough to capture a self-efficacy rating of 3. A similar objective of Pajares and Kranzler's (1995) study was to study calibration by examining students' overconfidence and underconfidence. Overconfidence was measured as students' self-efficacy ratings (4, 5 or 6 out of a possible 6) on items with incorrect solutions, while underconfidence was measured as ratings of 1, 2 or 3 with correct solutions. They found, overall, high school

students were overconfident in their capability for solving math problems. The researchers found no gender differences on calibration; however, they found ethnic differences: African-American students had lower math performance and lower self-efficacy ratings compared with their White counterparts. However, more African-American students were overconfident about their math capability despite their lower math performance.

Using different methods of measuring confidence judgments (Confidence Accuracy Quotient—CAQ), Lundeberg, Fox, and Puncochar (1994) examined undergraduate students' confidence after answering test questions. They found, overall, that both male and female students were overconfident, with males surprisingly overconfident when wrong. The gender differences of confidence and calibration were dependent on the content of the test items. In some content areas such as math and statistics, men were more confident than women; in other domains such as experimental design and the psychological sciences, no gender difference was found. Even though studies have measured calibration or accuracy judgments in many different ways, mean bias and mean accuracy were more appropriate, according to Schraw (1985).

Self-efficacy calibration was not the main objective studied by Pajares and Miller (1997) and Pajares and Graham (1999), but they did find that both measures of calibration (mean bias and mean accuracy) were significantly correlated with math performance. Pajares and Graham (1999) found that mean accuracy measure (fall, $\underline{r}$ = .88; spring, $\underline{r}$ = .81) and mean bias (fall, $\underline{r}$ = -.63; spring, $\underline{r}$ = -.68) correlated with math performance, exceeding the uncalibrated measure of self-efficacy or self-efficacy strength

(fall, r = .57; spring, r = .59) in magnitude. They found students were, overall, overconfident in their self-efficacy judgments; however, gifted students had less overconfidence than regular students. However, the data collection approach taken by Pajares and his colleagues assessing calibration had its limitations: namely, that students' math performance scores were also part of the mean bias and mean accuracy measures. Thus, the question of the value of self-efficacy calibration remains unanswered.

It is important to note that various studies have shown that students were generally overconfident and inaccurately calibrated in their math performances (Ewers & Wood, 1993; Lundeberg, Fox, & Puncochar, 1994; Pajares & Graham, 1999; Pajares & Kranzler, 1995; Pajares & Miller, 1997; Vermeer, Borkaerts, & Seegers, 2000). Further, it is important to understand why students are inaccurate in assessing their confidence judgments and performances. Moreover, some of the studies (Ewers & Wood, 1993; Lundeberg, Fox, & Puncochar, 1994; Vermeer, Borkaerts, & Seegers, 2000) found differences between male and female students' calibration. Thus, it is important to seek out the sources that possibly influence students' inaccurate self-efficacy calibrations.

Objectives and Hypotheses

Objective and Hypothesis 1. The first objective of the current study was to determine whether calibration measures of mathematics self-efficacy are separate dimensions of the self-efficacy construct. Calibration measures of mathematics self-efficacy were termed self-efficacy calibration bias and self-efficacy calibration accuracy (Pajares & Graham, 1999; Schraw, 1985). Self-efficacy calibration bias was

operationally defined as direction of judgment errors and self-efficacy calibration accuracy was defined as magnitude of judgment errors. Pajares and Graham (1999) found that these self-efficacy calibration measures had higher correlations than the uncalibrated self-efficacy measure (self-efficacy strength). Pajares and Graham (1999) found self-efficacy calibration bias and accuracy are significantly correlated in both data collection terms (fall, $r = -.39$; spring, $r = -.56$). In the present study, bias and accuracy correlated much higher ($r = -.87$), which indicated that the two calibration measures were assessing the same construct. To avoid colinearity, only one measure of self-efficacy calibration was used in the path analysis. Self-efficacy accuracy rather than bias was used in the path model because the former measured strength akin to self-efficacy strength. Further, self-efficacy accuracy is a positive value, unlike self-efficacy bias, which can be either a positive or negative value; when the mean was calculated, self-efficacy bias could deflate its predictability. Therefore, it was hypothesized that the self-efficacy calibration measure (accuracy) would increase the predictability of math performance above and beyond self-efficacy strength.

Objective and Hypothesis 2. The second objective of this study was to investigate the influence of mathematics self-efficacy strength and self-efficacy calibration accuracy on academic motivation such as effort. Self-efficacy has been shown to have a positive influence on students' motivation (Bong, 1997a; Multon, Brown, & Lent, 1993; Pajares & Graham, 1999; Pintrich et al., 1991, 1993). However, the issue of whether self-efficacy calibration influences students' motivation has not been tested to date. In addition, the present researcher asked students to judge effort after solving each math

problem and thus were item-specific judgments. It was hypothesized that self-efficacy and calibration accuracy measures would positively influence students' academic motivation such as effort.

Objective and Hypothesis 3. The third objective was to investigate possible gender differences on math self-efficacy, math self-efficacy calibration, motivation (effort), and math performance. Some evidence exists for gender differences in calibration. Ewers and Wood (1993) found boys to be more overconfident than girls, as did Vermeer, Borkaerts, and Seegers (2000), although Pajares and Kranzler (1995) did not find such results. Vermeer, Borkaerts, and Seegers also found girls to be more motivated when solving applied math tasks than boys were. Since the gender issue has been contested, further understanding and testing of gender on mathematics self-efficacy, mathematics self-efficacy calibration, motivation, and math performance are worth pursuing. It was thus hypothesized that, overall, students of both genders would likely be overcalibrated on their math self-efficacy, while males would be higher than females. Further, it was hypothesized that there would be no gender difference on math performance; however, females would be more likely to report more motivation (e.g., exert more effort) than boys.

Objective and Hypothesis 4. The fourth objective was to test whether students' previous math achievements influenced their math self-efficacy, math self-efficacy calibration, and math performance. According to Bandura (1996), students' previous mastery experiences provide them with information to gauge their capability for future events. Students' prior math achievement also provides an important statistical control in

the present study of math self-efficacy and math self-efficacy calibration. It was hypothesized that prior math achievement would have a positive influence on math self-efficacy, math self-efficacy calibration, and math performance.

Objective and Hypothesis 5. The fifth objective of this study was to investigate students' self-evaluation judgments of their math performance after solving math problems. Schraw et al. (1993) and Lundeberg et al. (1994) studied students' post-hoc performance accuracy judgments after solving tasks; however, no comparison has been conducted to date on self-efficacy judgments (before performance) and self-evaluation judgments (after performance). It was hypothesized that self-efficacy strength and self-efficacy calibration (accuracy) would be predictive of post-performance self-evaluation.

Objective and Hypothesis 6. The sixth objective of this study was to investigate students' math self-efficacy and math self-efficacy calibration in relation to levels of task difficulty (e.g., easy, moderate, and difficult). It was unclear why many students were poorly calibrated. Thus, it was important to examine whether task difficulty was an influence on students' math self-efficacy calibration. Although Schraw et al. (1993) found that students had higher performance and higher accuracy of judgments on the easy tasks, their measures were post-hoc, and thus are not generalizable to self-efficacy judgments, which are made prior to performance. It was hypothesized that with easy math problems, students would likely have higher math self-efficacy strength and be better calibrated on their math self-efficacy (accuracy).

Chapter III

METHODOLOGY

Participants

This study was conducted in May 2001; the participants were seventh graders recruited from four Catholic parochial schools in the state of Tennessee near the Nashville metropolitan area. A total of 163 seventh graders' parents were given letters of invitation and consent forms, asking them to permit their children to participate in the study. One hundred twenty-three (75.5%) parents returned the parental consent form and 121 (73.3%) signed and granted permission for their children to participate. Of the 121 children whose parents granted permission, 120 (72.7%) students signed the assent form and agreed to participate. The present study required students to participate in two sessions; those who were absent for a session were dropped from the data analysis. The final sample included in the data analysis was 107 students.

Instruments

Math Performance Test

The Math Performance Test items were adopted from the 1995 TIMSS (Third International Mathematics and Science Study) set of mathematics items specifically designed for seventh and eighth graders. For this particular set of math problems, TIMSS released 102 problems in multiple-choice, short-answer, and extended-response formats.

The item content represents the subject matter covered in schools in countries that participated in the 1995 TIMSS study, which also included schools in the United States. These items covered the content categories of (1) fractions and number sense, (2) algebra, (3) geometry, (4) data representation, analysis, and probability, and (5) proportionality. Further, the items were designed to meet performance expectations such as knowing, performing routine procedures, using complex procedures, and solving problems that are instituted in school mathematics. Even though the questions on the TIMSS were mostly multiple-choice, the present study adapted the questions to an open-ended format. Pajares and Miller (1997) studied both forms of assessments in math self-efficacy judgment and math performance. They concluded that an open-ended assessment format for math performance provides a more accurate measure of students' math capability and is less affected by guessing. Thus, an open-ended format increases the sensitivity of assessing calibration and math performance.

In order to choose the most appropriate items for the present study, selections were based on meeting the math curriculum guidelines provided by the Tennessee Catholic Archdiocese, and based on the content and difficulty codes of the TIMSS (Third International Mathematics and Science Study). The 15 math items covered all content areas pertaining to seventh grade mathematics: (1) fractions and number sense, (2) algebra, (3) measurement, (4) geometry, (5) data representation, analysis and probability, and (6) proportionality. In terms of performance expectation, 12 of the items involved solving problems and 3 involved performing routines and procedures. These 15 math problems were then presented in an open-ended format and as Form 1 of the math

performance tests. For this present study, a parallel form or Form 2 of the test was constructed. The 15 items in Form 2 were similar to math items in Form 1, with the modification of numerics to be computed. The main reason for keeping the two forms nearly identical was so that the math concepts, skills, and language of the items would not vary between forms. For example, Form 1, Question 5: *A class has 28 students. The ratio of girls to boys is* ***4:3****. How many girls are in the class?* and Form 2, Question 5: *A class has 28 students. The ratio of girls to boys is* ***5:2****. How many girls are in the class?* Because the math problems were modified into an open-ended format and a parallel form was constructed, these math performance tests were then pilot-studied before being administered to the participants in Tennessee.

Pilot Study. The main goals of the pilot study were to gather both reliability and validity evidence for the math tests. The purpose of the pilot study was fourfold: (1) to obtain alternate-form reliability, (2) to measure the research procedure time, (3) to test the modified math items from the TIMMS, and (4) to assure that the math tests were appropriate for the students. The pilot study was conducted with seventh graders in a Catholic parochial school in New York City in March of 2001. The school has two seventh-grade classes in which 35 students obtained parental consent and agreed to participate in the study. Of the 35 students, 32 showed up for both testing sessions. Students in one class were given Form 1 of the math test on the first day while the other class received Form 2. On the second day of testing, a reversed order of the test forms was administered.

Alternate-form reliability showed that the two math test forms had a correlation of equivalence of $r = .88$. The test procedures for both testing sessions were identical; the first testing session took an average of 40 minutes of class time and the second testing session took an average of 35 minutes. In the beginning of each testing session, the researcher read the directions and procedures verbatim, and asked the students if they understood the procedure and whether they needed any clarification. The researcher then went through examples with students in both sessions in order to ensure their understanding of the rating scales and familiarity with math problem layouts or formats. After students indicated they understood the purpose of the study and the procedures, the study continued.

For the math items, 15 paired t-tests between items in Form 1 and Form 2 were calculated. The results revealed that only two pairs of items showed statistical differences at $p < .05$, which indicated that students scored significant differences on those items between Forms 1 and 2. Those math items were then modified to improve their comparability for the main study conducted in Tennessee. To obtain their content validity evidence and appropriateness, the test items were selected on the basis of the TIMMS table of content specification. The table of content specification provides lists of seventh- and eighth-grade math content that are covered in most schools, including performance exceptions. Further, the TIMMS research office in Michigan provided the investigator with a chart of instructional time for each mathematical topic and concept taught in most seventh- and eighth-grade classrooms across the United States. Thus, the 15 math items selected for the pilot study proportionally represented the math content

areas covered in most middle schools, the performance expectations of students, and the instructional time devoted to those topics. In addition to gathering content validity evidence for the math items, in the first testing session while students were engaged in the study, the seventh-grade teacher was given a rating scale regarding the appropriateness of 30 math questions, which included the 15 items used in the pilot study. The teacher was asked two questions on each math problem: "How many of your students can solve this math problem?" with a scale ranging from 1 (none) to 4 (all), and "Is this an appropriate math problem for your students?" with a "yes" or "no" scale. If the teacher responded "no," she was instructed to explain her reasons. Of the 30 items, including the 15 on the pilot study, five were judged to be inappropriate; the teacher gave the same reason that those topics had yet to be covered in her class. Based on the teacher's rating, the math items used in the pilot study were appropriate for the grade level. At the conclusion of the second session, the researcher debriefed the participants of the study and answered any questions. As a result of the pilot study, Form 1 and Form 2 of the math performance tests were developed for the main study conducted in Tennessee.

Mathematics Self-efficacy Measure

The second instrument used in this study measured mathematics self-efficacy by asking students to indicate their confidence in solving each math question. Students then solved the questions for which they made self-efficacy judgments. This scale is a task-specific, corresponding assessment of students' perceptions of their capability to solve

each problem. According to Bandura (1997), such assessments of self-efficacy can increase predictions of academic performance. The math self-efficacy measure was worded as follows: "How confident are you about solving this math question correctly?" with a scale ranging from 1 (not at all) to 8 (completely).

Academic Motivation Measure—Effort Judgment

The third instrument measured students' post-performance effort judgment, which measured students' perceived effort exerted when solving a particular math problem. Bandura (1997) posited that students' self-efficacy influences their academic motivation; students who are efficacious are likely to persist longer and exert greater effort when encountering challenging tasks. Self-efficacy beliefs play an important mediating role in students' academic motivation which, in turn, influences academic achievement (Zimmerman, 2000). To measure effort judgment, students were asked to judge the amount of effort they exerted after they solved each math problem; this was a task-specific assessment of students' effort level. As Bong (1997a) suggested, the specificity of measures should be congruent between students' self-judgments and their performance. The effort judgment measure was worded as follows: "How much effort did you put in to solve this math problem?" with a scale ranging from 1 (none) to 8 (all).

Self-evaluation Measure

The fourth measure was the post-performance self-evaluation, which assesses, after problem solving, students' confidence in solving a problem correctly. Since Schraw

et al. (1993) and Lundeberg et al. (1993) asked participants to make accuracy judgments after the events, it was argued whether such a post-hoc process would provide information about performance and serve a self-evaluation purpose. In the present study, students were asked to make self-evaluation judgments on the math performance questions. The self-evaluation scale was specific and corresponded to each math problem. Accompanying each math problem was a self-evaluation measure worded as follows: "How confident are you that you solved this math problem correctly?" This measure had a scale ranging from 1 (not at all) to 8 (completely).

Previous Math Achievement

The fifth measure used in this study was students' previous math achievement scores. The participating schools used the ITBS (Iowa Tests of Basic Skills) as a measure of students' academic achievement. The students' previous (sixth grade) ITBS math scores in the forms of grade equivalence and national percentile were provided.

Procedures

Students participated in this study for two class periods. The between-test time interval was 3-4 days. The rationale for this time period between testing sessions was to prevent students from becoming "test-wise" since the two testing materials were similar (Gall, Borg, & Gall, 1996). Pajares and Graham (1999) found a high correlation between calibration measures and math performance; however, students' math performance scores were also part of the calibration measures. Thus, to overcome one limitation of Pajares

and Graham's (1999) study, two testing sessions were necessary. The data gathered from Session 1 were used to calculate calibration accuracy and bias, while the data gathered from Session 2 were used to obtain students' self-efficacy, post-performance self-evaluation, and effort judgment as well as the dependent measure, math performance. In this study, measures were administered in each math class as a group.

In Session 1, students were given Form 1 (see Appendix A), which contains (1) an answer sheet on which students circled their math self-efficacy responses; (2) an example; (3) 15 math problems with an effort judgment scale and a self-evaluation scale accompanying each math problem; and (4) three general persistence questions on the last page. First, the researcher used an overhead projector to show students a math question as an example; students were instructed to read the example on the screen and make a self-efficacy judgment without solving the example on their answer sheets. Once two examples were given and students understood the procedures, the researcher presented the 15 math questions one at a time. As each math problem was shown on the screen, the researcher also read the problem aloud to students. Once a question was read aloud, the researcher covered that question and instructed students to circle a rating of their perceived math self-efficacy corresponding to the particular math problem. The brief exposure to each question allowed students to assess the difficulty of the question without having time to actually solve it. After the researcher went through the 15 math problems and students made their self-efficacy judgments, students were informed they would have a chance to solve the math problems. Second, students were shown a math example on the screen, which had the same layout as the actual 15 math problems. The researcher

reviewed one math example with the students and instructed them to rate their effort and self-evaluation scales located at the bottom of the page. Accompanying each math problem was an effort judgment scale to assess students' perceived effort exerted when solving the math problem. In addition, students were asked to rate a self-evaluation scale to assess their performance on that particular math problem. Third, once students understood what they were asked to do with the math example, they began to solve the actual 15 math problems and rate the accompanying scales. Students were given one class period to solve the problems and rate the scales accompanying each. This concluded Session 1.

In Session 2, students were given Form 2 (see Appendix B). The procedures of Session 2 were exactly the same as those of Session 1. The only difference was that the math problems contained different numerals from those in Session 1. First, the researcher showed each math problem on the screen and students were asked to make self-efficacy judgments about each math question without solving it. Second, the researcher went over the example page with students to familiarize them with the layout of the math problems and the accompanying effort judgment and self-evaluation scales. Third, students were instructed to solve each math problem, to rate the amount of effort involved in solving each math problem, and to make self-evaluations on each math problem. This concluded Session 2. Since Session 2 had same procedures as Session 1, students were able to complete this session in less time. After completing Session 2, each participant was given a pencil as a token of appreciation. The researcher provided a copy

of the math questions and answers to the participating teachers who, in turn, made the questions and answers available to all students.

Each student was identified by an assigned case number. During the collection of the participants' previous math achievement scores, only the authorized school personnel and the researcher knew the participants' names and corresponding numbers. Once the participants' demographic information, such as gender, ethnicity, and previous math achievement scores, was recorded, participants' names were removed. Students' case numbers were then used to link to the data collected during the two testing sessions for data analysis.

Data Analysis

First, demographic information, which was gathered from school records, was categorized. As for ethnicity, the students were 98% Caucasian American, and as a result, ethnicity was not included in data analysis.

Second, calibration of self-efficacy was calculated using data gathered from Session 1. Even though only the calibration accuracy was used in the path analysis portion of the data analysis, both measures (bias and accuracy) of self-efficacy calibration were computed. Mean bias and mean accuracy were computed for each participant following the procedures suggested by Schraw (1985) and Schraw, Potenza, and Bebelsick-Gullet (1993). To compute self-efficacy calibration bias, each correct answer was scored an 8 and each incorrect score a 1. These 1 or 8 scores corresponded to the self-efficacy scores ranging from 1 to 8. Thus, a student who expressed "not at all

confident" (1) in solving the problem correctly but indeed missed the problem (1) indicated a zero bias (1 minus 1). On the other hand, a student with the same lack of confidence who actually correctly solved the problem received a bias score of -7 (1 minus 8), indicating underconfidence. Thus, self-efficacy calibration bias scores could range from -7 to +7. Self-efficacy calibration accuracy scores were computed by subtracting the absolute value of each bias score from 7, indicating the magnitude of the judgment error, with a range of 0 (complete inaccuracy) to 7 (complete accuracy).

Third, from Session 2, item difficulty was computed for all 15 math problems used in the study. The item difficulty in percentage was equal to the number of students who solved the item correctly divided by the number of students who tried the item. The higher the percentages, the easier the items; the lower the percentages, the more difficult the items. The 15 math items were then ranked according to the percentage of participants answering correctly. Easy items were those which were answered correctly by approximately 65% or more of students, while difficult items were those which were answered correctly by less than 30% of students. The rest of the items were considered to be moderately difficult items.

Fourth, also from Session 2, each student's self-efficacy ratings, which accompanied the 15 math items, were averaged in order to calculate self-efficacy. The same calculation procedure was also used to obtain post-performance self-evaluation and effort judgment scores. Fifth, from Session 2, each student's math performance scores (dependent variable) were collected.

Descriptive and correlational statistical analyses were conducted in this study. Descriptive statistics of self-efficacy strength, self-efficacy calibration bias, self-efficacy calibration accuracy, motivation measure (effort judgment), self-evaluation, math performances, and previous math achievement were presented according to the category of gender. A General Linear Model (GLM)—Repeated-Measure Design—was performed to test any significance between gender and within item difficulty level (3 levels) on the five measures: self-efficacy strength, self-efficacy calibration bias, self-efficacy calibration accuracy, effort judgment, and self-evaluation. The "repeated measure" option was used in analyzing the data because all participants were exposed to the same math problems of varying item difficulty and the same measures that accompanied those math problems. Following the GLM (repeated measures), post-hoc analyses were performed to further identify specific mean differences among item difficulty levels. Correlations were calculated to examine associations among the following measures: self-efficacy calibration bias, self-efficacy calibration accuracy, self-efficacy strength, effort judgment, self-evaluation, math performance, previous math achievement scores, and gender. A linear regression was also performed to assess the independent contribution of calibration accuracy beyond the uncalibrated self-efficacy in predicting math performance. Finally, using LISREL 8 (Joreskog & Sorbom, 1996), path analysis was performed to test the influence of calibration accuracy, self-efficacy, and previous math achievement (ITBS) on math performance, post-performance self-evaluation, and effort judgment.

To answer Objectives 1, 2, 4, and 5, a path analysis was conducted to test the direct and indirect effects among variables: self-efficacy calibration accuracy, self-efficacy strength, gender, previous math achievement (ITBS), math performance, post-performance self-evaluation, and effort judgment. In addition, a regression analysis was conducted to answer Objective 1.

Objective 1—To determine whether self-efficacy calibration measure (accuracy) was an important dimension of self-efficacy construct.

Objective 2—To investigate the influence of mathematics self-efficacy and self-efficacy calibration accuracy on academic motivation such as post-performance effort judgment.

Objective 4—To test whether students' prior math achievement influenced their math self-efficacy, calibration, and math performance.

Objective 5—To examine students' post-performance self-evaluation judgment.

To answer Objectives 3 and 6, a GLM with repeated measures was conducted to examine the effect of gender and level of item difficulty on the following dependent variables: self-efficacy calibration bias, self-efficacy calibration accuracy, self-efficacy strength, post-performance self-evaluation, and effort judgment. Post-hoc analysis, Bonferroni, was performed to identify specific differences among three levels of items.

Objective 3—To investigate possible gender differences on math self-efficacy, calibration measures, post-performance effort judgment and self-evaluation.

Objective 6—To examine students' math self-efficacy and calibration measures in relation to levels of task difficulty (e.g., easy, moderate, and difficult).

Chapter IV

RESULTS

Reliability of Test Forms

First, the results showed that the correlation between the two Forms was .84. Second, the split-half reliability was computed to determine the equivalence between two parallel forms. The Guttman split-half reliability coefficient of the overall test was .91. Since both Forms had the same number of items, both equal-length Spearman-Brown and unequal-length Spearman-Brown reliabilities were also .91. Third, the internal consistency of all 30 items was $\alpha = .88$. Split-half reliability (Guttman split-half = .91) showed that the two Forms were highly equivalent and the internal consistency of each Form was nearly identical (Form 1, $\alpha = .784$; Form 2, $\alpha = .780$), indicating that the means and variances in each scale did not differ much (George & Mallery, 2001).

Difficulty Level of the Test Items

Since students' math performance was based on Form 2, item difficulty (percentage of students answering the item correctly over all those attempting the item) was computed for each math question. Math items were categorized into three levels: easy, moderate, and difficult. Based on Form 2 questions, the easy items were questions 1, 2, 8, and 12, with about 65% or more students solving the items correctly. The moderately difficult items were questions 3, 5, 7, 9, 13, and 15, while the difficult items were questions 4, 6, 10, 11, and 14, with less than 30% of the students answering

correctly. The math questions were constructed with various item difficulty levels in accordance with TIMSS results and the results from the pilot study. Thus, the rationale for the 65% and 30% cutoffs was to separate the items into three approximately-equal groups (4 easy, 6 moderate, and 5 difficult questions).

Role of Gender and Item Difficulty on Dependent Measures

This study had six dependent measures: math performance, self-efficacy calibration accuracy and bias, self-efficacy, and post-performance self-evaluation and effort judgment. As a group, students scored approximately 7 out of 15 questions correctly (M = 6.64, SD = 3.29), with girls (M = 6.28, SD = 3.22) scoring slightly lower than boys (M = 7.19, SD = 3.36). ANOVA was computed to compare means between the genders in the dependent variables. No statistical difference was found on math performance between genders, F (1, 105) = 1.99, p = .16. For other dependent measures, a breakdown by gender and three levels of item difficulty (easy—Level 1, moderate—Level 2, difficult—Level 3) is shown in Table 1.

The General Linear Model (GLM) with a repeated measure design was performed to test how gender and item difficulty levels influenced students' self-efficacy strength, calibration accuracy, calibration bias, and post-performance self-evaluation and effort judgment. Item difficulty was a within-subject measure while gender was a between-subject measure. Nested under the three levels of item difficulty were repeated measures of self-efficacy (e.g., mean self-efficacy for Level 1 items, mean self-efficacy for Level 2

Table 1

Means and Standard Deviations for Calibration Accuracy, Bias, Self-efficacy, Effort Judgment, and Self-evaluation as a Function of Gender and Level of Item Difficulty

			Item Difficulty Levels							
			Easy Level 1		Moderate Level 2		Difficult Level 3		Combined	
Group	Measures[a]	n	M	SD	M	SD	M	SD	M	SD
Girls		65								
	Accuracy[b]		5.14	1.56	3.56	1.29	2.97	1.61	3.78	0.97
	Bias[c]		0.99	1.66	2.42	1.93	3.48	2.02	2.46	1.28
	Self-efficacy		7.66	0.72	6.06	1.29	5.94	1.42	6.45	1.07
	Effort judgment		2.31	1.35	3.01	1.44	3.10	1.66	2.85	1.39
	Self-evaluation		7.49	0.80	5.88	1.47	5.63	1.65	6.23	1.20
Boys		42								
	Accuracy		5.01	1.42	3.54	1.39	2.96	1.45	3.74	1.17
	Bias		1.07	1.77	2.25	2.09	3.51	1.92	2.39	1.55
	Self-efficacy		7.68	0.61	6.78	1.28	6.33	1.60	6.87	1.12
	Effort judgment		2.15	0.76	2.39	1.00	2.50	1.30	2.36	0.91
	Self-evaluation		7.63	0.58	6.63	1.48	6.17	1.75	6.75	1.22

Note. N = 107.

[a]These measures (self-efficacy, effort judgment, and self-evaluation) ranged from 1 to 8, with 8 being the highest possible number.

[b]Accuracy ranged from 0 through 7, with 7 being the highest accuracy.

[c]Bias ranged from -7 to +7, with negative numbers indicating underconfidence and positive numbers indicating overconfidence.

items, and mean self-efficacy for Level 3 items). Similarly, nested under the three levels of item difficulty were repeated measures of calibration accuracy, calibration bias, effort judgment, and self-evaluation.

An overall multivariate test of significance showed no gender effect on the array of dependent variables: calibration accuracy, bias, self-efficacy, effort judgment, and self-evaluation, Wilks' lambda = .94, $\underline{F}$ (5, 101) = 1.30, $\underline{p} > .05$. There was a significant item difficulty level effect on the dependent measures: calibration accuracy, bias, self-efficacy, effort judgment, and self-evaluation, Wilks' lambda = .26, $\underline{F}$ (10, 412) = 40.17, $\underline{p} < .01$. Univariate tests of significance showed an item difficulty level effect for each dependent variable: calibration accuracy, $\underline{F}$ (2, 210) = 74.69, $\underline{p} < .01$; bias, $\underline{F}$ (2, 210) = 61.76, $\underline{p} < .01$; self-efficacy, $\underline{F}$ (2, 210) = 128.66, $\underline{p} < .01$; effort judgment, $\underline{F}$ (2, 210) = 19.27, $\underline{p} < .01$; and self-evaluation, $\underline{F}$ (2, 210) = 100.81, $\underline{p} < .01$.

Bonferroni multiple-comparison analyses were performed to compare dependent measure means among three levels of item difficulty (easy, moderate, and difficult) (see Table 2). With regard to calibration accuracy, students were significantly more accurate on easy items than on moderate items and, in turn, were significantly more accurate on moderate items than on difficult items. Students were significantly more biased on difficult items than on moderate items and, in turn, were significantly more biased on moderate items than on easy items. As for post-performance self-evaluation, students were significantly more confident of their performance on easy items than on moderate items and, in turn, were significantly more confident on moderate items than on difficult

items. Students were significantly higher in their math self-efficacy on easy items than on moderate items and, in turn, were significantly higher on moderate items than on

Table 2

Post-hoc Analysis on Dependent Measures of Calibration Accuracy, Bias, Self-efficacy, Effort Judgment, and Self-evaluation as a Function of Level of Item Difficulty

Measures	Item Difficulty Levels: Easy Level (1)[a] M	Easy Level (1)[a] SD	Moderate Level (2)[a] M	Moderate Level (2)[a] SD	Difficult Level (3)[a] M	Difficult Level (3)[a] SD	Multiple Comparisons[b]
Accuracy	5.09	1.50	3.55	1.32	2.97	1.54	1>2>3
Bias	1.02	1.70	2.35	1.99	3.50	1.97	1<2<3
Self-efficacy	7.67	0.67	6.34	1.33	6.09	1.50	1>2 >3
Effort Judgment	2.25	1.15	2.77	1.32	2.86	1.55	1<2, 1<3, 2=3
Self-evaluation	7.55	0.72	6.17	1.51	5.84	1.70	1>2>3

Note.
[a]The numbers are used to illustrate significant differences in the last column titled "Multiple Comparisons."
[b]Bonferroni, $p < .01$

difficult items. Interestingly, with regard to effort judgment, students judged that significantly less effort was put forth on easy items than on moderate items, but effort judgment was not significantly different on difficult items ($p = .63$).

Polynomial contrasts were further conducted to test the trends of the dependent measures as item difficulty increased. The analyses showed that all five dependent measures (calibration accuracy, bias, self-efficacy, effort judgment, and self-evaluation) had significant linear trends as item difficulty levels increased ($p < .01$). In other words, for each dependent measure, the means increased proportionately and significantly as the items became more difficult.

Influence of Calibration Accuracy and Item Difficulty Levels on Self-efficacy, Post-performance Self-evaluation, Effort Judgment, and Math Performance

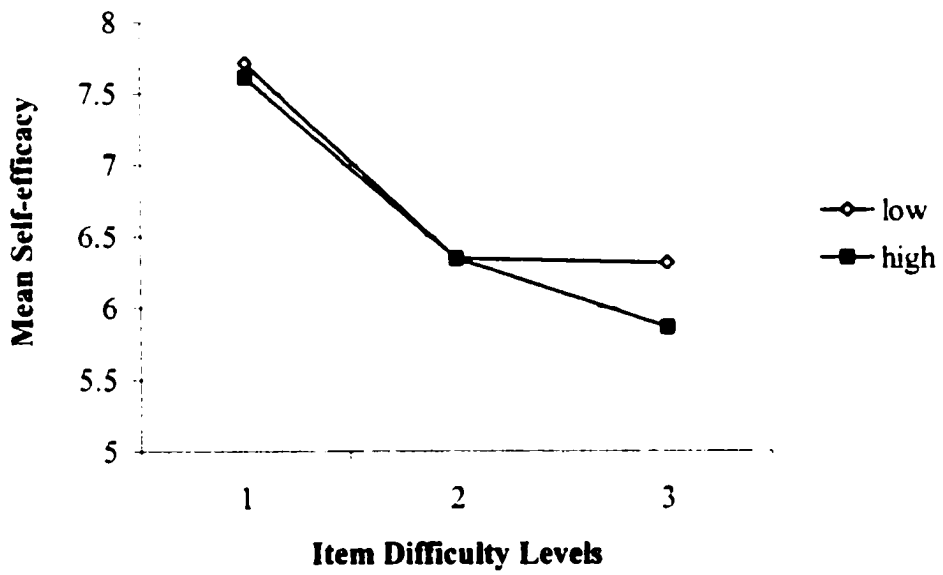
Separate analyses were conducted to test how students with different calibration accuracy levels rated on their self-efficacy, and post-performance self-evaluation and effort judgment. The calibration accuracy scores could range from 0 (complete inaccuracy) to 7 (complete accuracy). Students were divided into two groups based on the calibration accuracy scores midpoint of the scale. Students who had average calibration accuracy scores below "4" were in the low calibration group, and those whose scores were higher than or equal to "4" were in the high calibration group. Of all the participants ($N = 107$), 51 were in the high calibration group and 56 were in the low calibration group.

Three separate GLM repeated-measure designs were conducted for self-efficacy, self-evaluation, and effort judgment, with calibration groups (high, low) as a between-subject factor and item difficulty levels (easy, moderate, difficult) as a within-subject factor. For the self-efficacy measure, there was no significant calibration effect, F (1, 105) = .83, $p > .05$. The item difficulty main effect was also significant, F (2, 210) = 141.28, $p < .01$, and interaction between calibration and item difficulty levels nearly attained statistical significance, F (2, 210) = 2.79, $p = .06$. Figure 1 depicts the interaction between calibration and item difficulty levels on self-efficacy. Both low (M = 7.71, SD = .09) and high calibration (M = 7.61, SD = .09) groups reported very similar self-efficacy with easy items, but as items became moderately difficult, both low calibration students (M = 6.34, SD = .18) and high calibration students (M= 6.34, SD = .19) dropped their self-efficacy to the same level. Interestingly, as items became difficult, the self-efficacy of low calibration students (M = 6.31, SD = .20) dropped only very slightly, compared to the high calibration group (M = 5.86, SD = .21).

As for effort judgment, the performance of the two calibration groups was statistically comparable, F (1, 105) = 3.11, $p = .08$. However, there was a significant item difficulty effect on effort judgment, F (2, 210) = 23.02, $p < .01$. No interaction was found between calibration and item difficulty levels, F (2, 210) = .15, $p > .05$. In other words, as item difficulty levels increased, both calibration groups increased their effort when solving the problems.

Figure 1

Means of Self-efficacy of High- and Low-Calibrated Students according to Item Difficulty Levels



For the post-performance self-evaluation measure, the performance of two calibration groups was statistically comparable, F (1, 105) = .20, $p > .05$. However, there was a significant item difficulty effect, F (2, 210) = 111.54, $p < .01$, as well as a significant interaction between calibration groups and item difficulty, F (2, 210) = 3.30, $p < .05$ (see Figure 2). The low-calibrated students (M = 7.62, SD = .10) had a slightly higher mean on self-evaluation than the high-calibrated group (M = 7.47, SD = .10) on easy items. Both groups dropped their post-performance self-evaluation on moderate items, with the low-calibrated group (M = 5.96, SD = .20) doing so more than the high-calibrated group (M = 6.41, SD = .21). Both groups had very similar post-performance self-evaluations on difficult items (low group, M = 5.84, SD = .23; high group, M = 5.85, SD = .24).

For math performance, the scores of the high and low calibration groups differed significantly, F (1, 105) = 40.85, $p < .01$. There was a significant item difficulty effect, F (2, 210) = 130.77, $p < .01$, as well as a significant interaction between calibration groups and item difficulty, F (2, 210) = 16.46, $p < .01$ (see Figure 3). With easy items, the low-calibrated group scored almost three items (M = 2.89, SD = .12), while the high-calibrated group scored more than three items (M = 3.28, SD = .13). With moderate difficulty items, the low-calibrated group scored much lower (M = 1.64, SD = .21) than the high-calibrated group (M = 3.53, SD = .22). Again, with difficult items, the low-calibrated group (M = .45, SD = .15) scored lower than the high-calibrated group (M = 1.65, SD = .16).

Figure 2

Means of Post-performance Self-evaluation of High- and Low-Calibrated Students according to Item Difficulty Levels

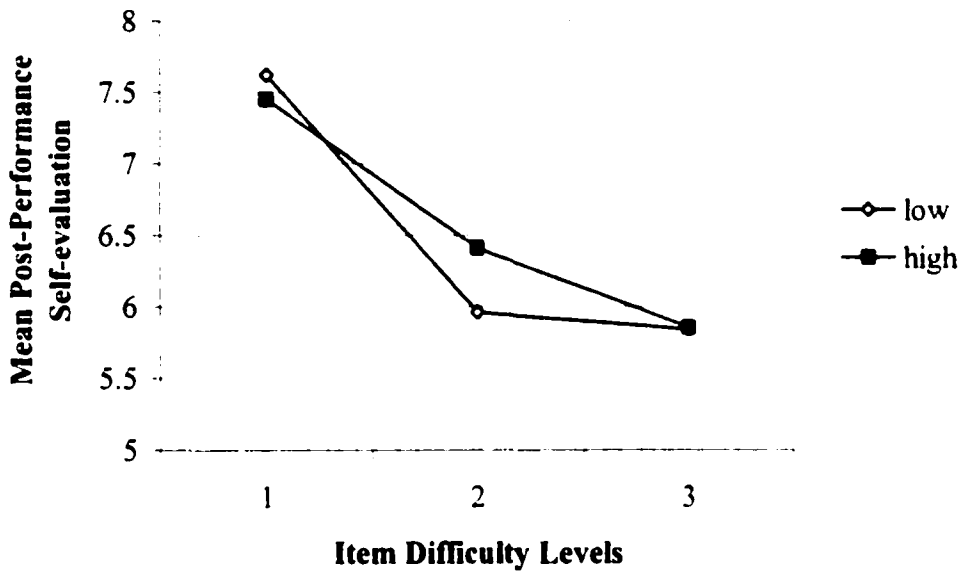
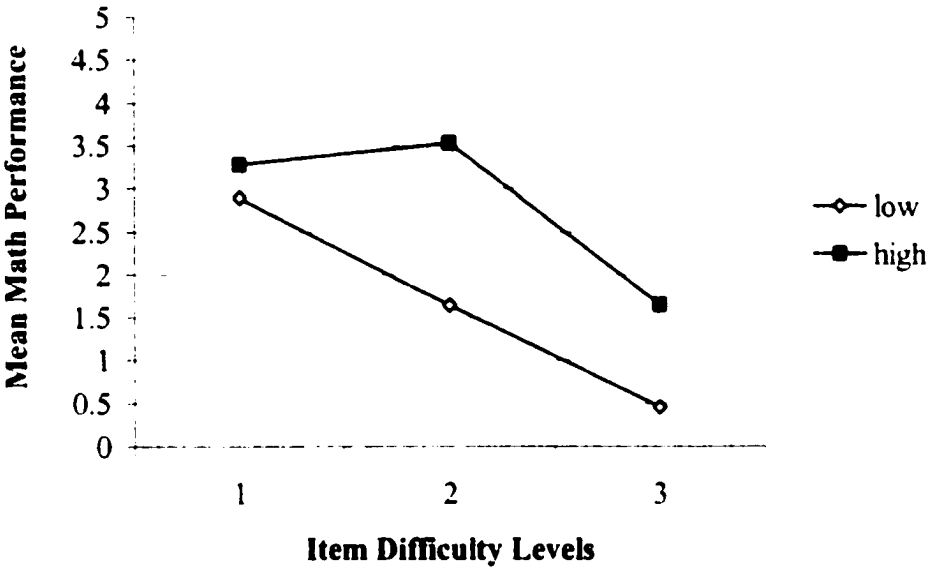


Figure 3

Means of Math Performance of High- and Low-Calibrated Students according to Item Difficulty Levels



Correlations

Pearson correlation coefficients for the measures are reported in Table 3. The following significant correlations were of particular interest for the present: math performance with calibration accuracy (r = .63), self-efficacy (r = .50), and prior math achievement (r = .65). Self-efficacy was highly, yet negatively, correlated with the motivational variable, post-performance effort judgment (r = -.49). Post-performance self-evaluation was highly and positively correlated with self-efficacy (r = .77).

Table 3

Correlations among Measures

Measures	2	3	4	5	6	7	8
1. Math Performance	-.62**	.63**	.50**	-.28**	.50**	.65**	-.14
2. Bias	-----	-.87**	.05	.12	-.10	-.40**	.03
3. Accuracy		-----	-.01	-.16	.09	.44**	.02
4. Self-efficacy			-----	-.49**	.77**	.42**	-.19
5. Effort Judgment				-----	-.47**	-.26**	.19*
6. Self-evaluation					-----	.37**	-.21**
7. ITBS[a]						-----	-.12
8. Gender[b]							-----

Note.
[a]Iowa Test of Basic Skills (national percentile scores: males, M = 68.90, SD = 25.67; females, M = 65.57, SD = 21.35)
[b]Coding (male = 1; female = 2)
* $p < .05$
** $p < .01$

Regression Analysis

The purpose of the regression analysis was to discover if calibration (accuracy) added to an improved the prediction of the uncalibrated self-efficacy measure. Math performance was the dependent measure, while self-efficacy and calibration accuracy were the independent variables. Self-efficacy first entered the equation, followed by calibration accuracy. As expected, self-efficacy (β = .51) and calibration accuracy (β = .63) significantly predicted math performance. Both variables accounted for about 65% of the variance in math performance, with self-efficacy accounting for 25.4% and calibration accuracy accounting for 40%. The significant change in R^2 revealed that calibration accuracy made an important and independent contribution to predicting math performance beyond the uncalibrated self-efficacy, $\underline{R}^2$ change = .40, $\underline{F}$ (1, 104) = 117.98, $\underline{p} < .01$.

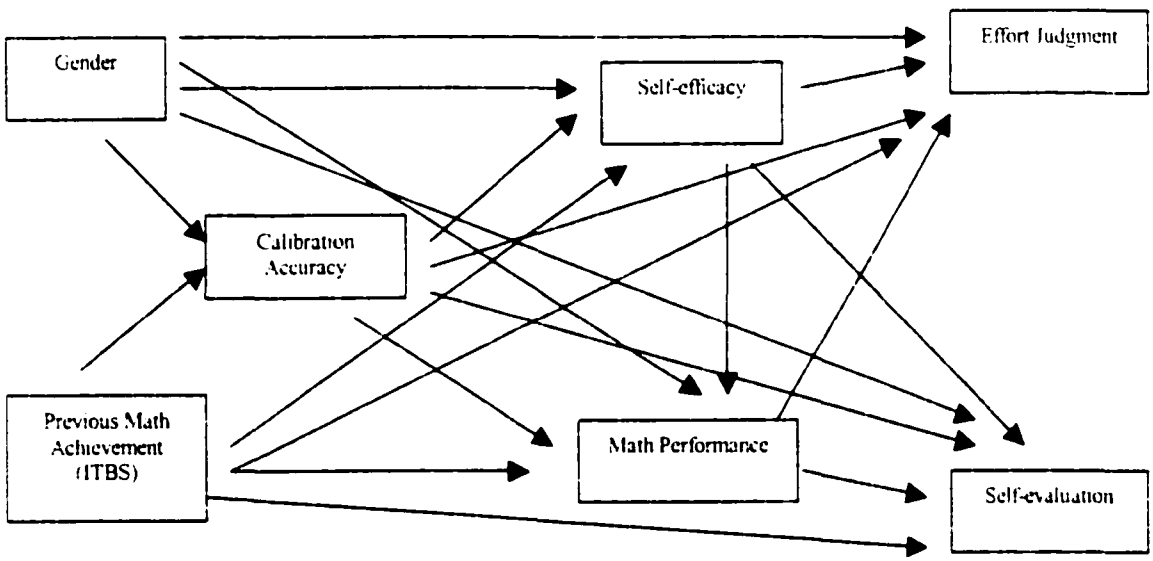
Path Analysis

To test the implied causal role of students' calibration (accuracy) on their self-efficacy, math performance, and post-performance self-evaluation and effort judgment, a path model was proposed (see Figure 4). In addition, gender was included in the model to test whether it had a direct effect on predicting calibration accuracy, self-efficacy, math performance, and post-performance self-evaluation and effort judgment. Prior math achievement was also included in the model to test its direct effect on calibration accuracy, self-efficacy, math performance, and post-performance self-evaluation and effort judgment. Historically, previous performance on a measure is the best predictor of

subsequent performance, and the relative influence of previous performance in the path model provides a useful standard of comparison for the other variables. Because of the extremely high correlation between calibration bias and accuracy, and the greater ease of interpreting the latter measure, only calibration accuracy was included in the path analysis.

Figure 4

Proposed Path Model



The proposed model was evaluated using LISERL 8 (Joreskog & Sorbom, 1996). Estimation of the proposed path model revealed a nonsignificant χ^2 value, χ^2 (1, N = 107) = 2.59, p = .11; a GFI adjusted for degree of freedom (AGFI) of .81; and a comparative fit index (CFI) of .99. Although all indices showed a reasonable goodness of fit, nine of

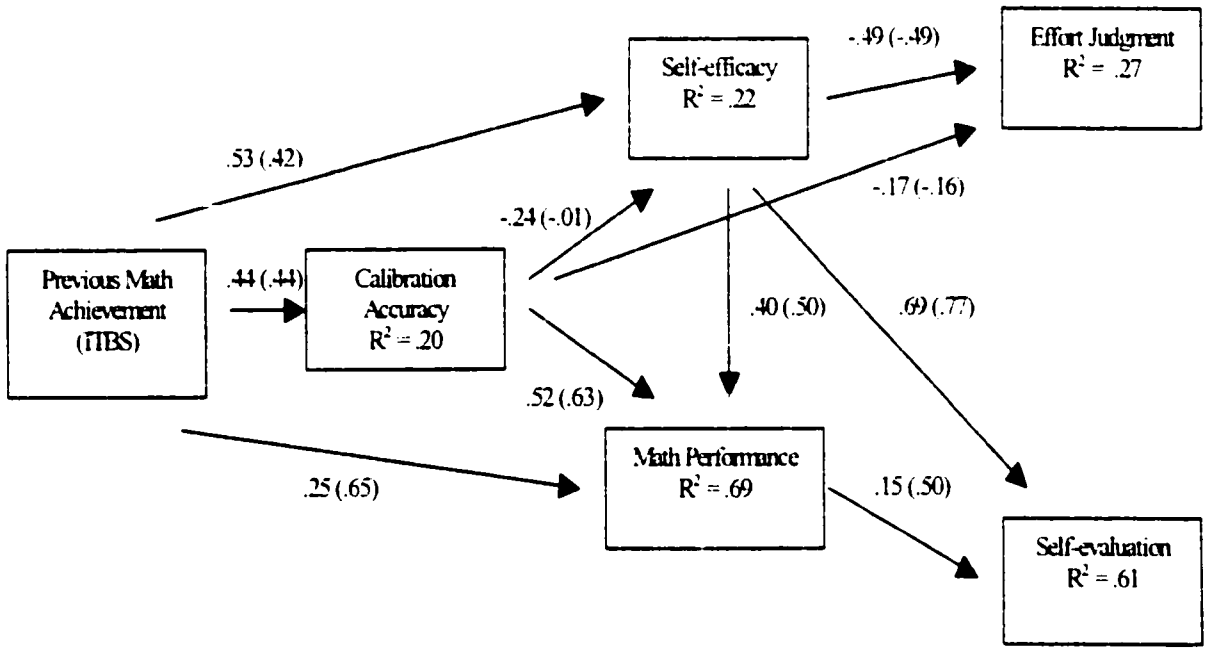
the paths were nonsignificant: (1) from calibration accuracy to post-performance self-evaluation, (2) from gender to calibration accuracy, (3) from gender to self-efficacy, (4) from gender to effort judgment, (5) from gender to post-performance self-evaluation, (6) from gender to math performance, (7) from previous math achievement to effort judgment, (8) from previous math achievement to self-evaluation, and (9) from math performance to effort judgment. After removing those nonsignificant paths, the final model (see Figure 5) again revealed a nonsignificant χ^2 (5, N = 107) = 5.10, p = .40; a GFI adjusted for degree of freedom (AGFI) of .93; and a comparative fit index (CFI) of 1.00.

Figure 5 presents the significant paths (standardized coefficients) resulting from the LISERL analysis. Standardized path-coefficient β's are denoted outside parentheses, while Pearson-Product correlation coefficients are denoted inside parentheses. As expected, students' calibration accuracy (β = .52), self-efficacy (β = .40), and prior math achievement (β = .25) had direct effects on math performance. Further, students' math performance (β = .15) and self-efficacy (β = .69) had direct effects on post-performance self-evaluation. Interestingly, students' self-efficacy (β = -.49) and calibration accuracy (β = -.17) had negative direct effects on their effort judgment. Also worth noting was that calibration accuracy (β = -.24) had a negative direct effect, while prior achievement (β = .53) had a positive direct effect on self-efficacy. Prior achievement also had a positive direct effect (β= .44) on calibration accuracy. Table 4 depicts the decomposition of effects from the path analysis.

The predictor variables (self-efficacy, calibration accuracy, and prior math performance) accounted for 69% of the variance in math performance, while 61% of the variance in the post-performance self-evaluation was explained by all measures in the path model. Further, calibration accuracy and prior math achievement accounted for 22% of the variance in self-efficacy, and 27% of the variance in students' effort judgment. Prior math achievement accounted for 20% of the variance in students' calibration accuracy.

Figure 5

Final Reduced Path Model



Note. All path coefficients are significant at .05 level.

Table 4

Decomposition of Effects from the Path Analysis

Effect	Standardized estimate (β)	Standard error (SE)	t[b]	R^2
On accuracy				.20
of ITBS[a]	.44	.087	5.06	
On self-efficacy				.22
of accuracy	-.24	.096	-2.48	
of ITBS	.53	.096	5.49	
On effort judgment				.27
of accuracy	-.17	.083	-1.99	
of self-efficacy	-.49	.083	-5.92	
On self-evaluation				.61
of self-efficacy	.69	.071	9.73	
of math performance	.15	.071	2.14	
On math performance				.69
of accuracy	.52	063	8.27	
of self-efficacy	.40	.062	6.49	
of ITBS	.25	.069	3.65	

Note.
[a]ITBS (Iowa Test of Basic Skills) previous math achievement (N = 107)
[b]$p < .05$

Previous math achievement (ITBS) had a significant indirect effect on post-performance self-evaluation, also mediated by self-efficacy (β = .39). Self-efficacy had both direct and indirect effects (β = .06) on post-performance self-evaluation. Calibration accuracy had a significant positive direct effect yet a significant negative indirect effect on math performance, which was mediated by self-efficacy (β = -.10). As for previous math achievement, it had a significant direct effect, and a large indirect effect (β = .40) on math performance via self-efficacy. Calibration accuracy had a negative direct effect, yet a positive indirect effect (β = .12) on effort judgment mediated through self-efficacy. Interestingly, the total effect of calibration accuracy on effort judgment (β = -.05) did not reach statistical significance ($t(107) = -.49$, $p > .05$). Previous math achievement had a significant negative indirect effect (β = -.28) on effort judgment mediated by self-efficacy. Previous math achievement had a significant direct effect and an indirect effect (β = -.11) on self-efficacy mediated through calibration accuracy. As the path analysis shows, self-efficacy and calibration accuracy are key mediating variables affecting dependent measures such as math performance, effort judgment, and post-performance self-evaluation. Table 5 depicts all direct, indirect, and total effects of the final path model.

Table 5

Direct, Indirect, and Total Effects on Math Performance, Self-evaluation, Effort Judgment, and Self-efficacy

Effect	Direct effect	Indirect effect	Total effect
On accuracy			
of ITBS[a]	.44	.00	.44
On math performance			
of accuracy	.52	-.10	.42
of ITBS	.25	.40	.65
of self-efficacy	.40	.00	.40
On self-evaluation			
of accuracy	.00	-.10[b]	-.10[b]
of ITBS	.00	.39	.39
of self-efficacy	.69	.06	.75
of math performance	.15	.00	.15
On effort judgment			
of accuracy	-.17	.12	-.05[b]
of self-efficacy	-.49	.00	-.49
of ITBS	.00	-.28	-.28
On self-efficacy			
of accuracy	-.24	.00	-.24
of ITBS	.53	-.11	.42

Note. N = 107

[a]ITBS (Iowa Test of Basic Skills) indicates students' previous math achievement.

[b]These effects are not statistically significant at the $p > .05$ level.

Chapter V

DISCUSSION

General Discussion

The primary aim of the present study was to test whether the math self-efficacy calibration of seventh graders was an important dimension of the self-efficacy construct by examining its role as a predictor of math performance and post-performance self-evaluation and effort judgments. In addition, the present study explored the sources that contributed to students' self-efficacy calibration. To answer this question, high- and low-calibrated students were compared across three levels of task difficulty. Finally, individual differences in students' gender and previous math achievement were taken into account. To answer these questions, a path analysis was used to test a model based on social-cognitive theory (Bandura, 1986, 1997), prior research on self-efficacy (Ewers & Wood, 1993; Hackett & Betz, 1989; Pajares & Graham, 1999; Pajares & Kranzler, 1995; Pajares & Miller, 1994), issues of gender and math (Eccles, 1994; Hyde, Fennema, & Lamon, 1990; Hyde et al., 1990), motivational factor (effort judgment) (Bong, 1997a; Pintrich et al., 1991, 1993; Schunk, 1983, 1991; Vermeer et al., 2000), and calibration (Lundeberg et al., 1994; Schraw, 1995; Schraw et al., 1993). Further analysis using the General Linear Model repeated-measure designs was conducted to answer questions regarding task difficulty levels as well as differences between calibration groups.

Findings from the path analysis revealed that the proposed model fit the obtained data very well (CFI = 1.00), and accounted for 69% of the variance on math performance

as well as 61% of the variance on post-performance self-evaluation. Most important, the path model indicated that self-efficacy calibration accuracy was empirically found to be an important dimension of self-efficacy. Calibration accuracy had contributed to predicting and mediating math performance, and post-performance self-evaluation and effort judgment. As for self-efficacy, it was also found to have both predictive and mediational effects on math performance, and post-performance self-evaluation and effort judgment. Clearly, the findings of this study showed that math self-efficacy was a crucial variable in predicting students' math performance, and post-performance self-evaluation and effort judgment. As in other studies (Pajares & Kranzler, 1995; Pajares & Miller, 1994), self-efficacy also played an important mediational role. It mediated both calibration accuracy and previous achievement in predicting math performance and post-performance self-evaluation. As for how students varied in their calibration, the present study found that high-calibrated students were significantly more sensitive to the various levels of task difficulty, compared to low-calibrated students. The following section discusses in detail each variable tested in this study and addresses each hypothesis presented here.

Hypotheses

Hypothesis 1

The first hypothesis asked whether self-efficacy calibration was an important dimension of the self-efficacy construct that contributes to its predictive power. Following Bandura's recommendations, the measures of self-efficacy in the present study

assessed differences in self-efficacy strength (from no to complete confidence) and level (three difficulties of items), but not differences in generality (to other domains of tasks). Bandura (1997) had, however, discussed the role of calibration in self-efficacy by cautioning that an individual should have reasonably matched self-efficacy judgments and actions. The present findings showed, in fact, that self-efficacy calibration (accuracy) improved the predictiveness of students' math performance above and beyond uncalibrated measures. A regression analysis showed that calibration accuracy significantly and independently explained nearly 40% of the variances in predicting math performance beyond the uncalibrated self-efficacy. In addition, path analysis showed that calibration accuracy had a direct influence on predicting students' math performance as well as an indirect influence via self-efficacy. Interestingly, the mediating role of self-efficacy minimized some of the direct effects of calibration accuracy on math performance due to the negative relationship between calibration accuracy and self-efficacy (see Table 5 for effects of accuracy on math performance). As much self-efficacy research has indicated, a higher sense of self-efficacy positively influences student achievement (Multon, Brown, & Lent, 1991); however, the present study indicated that a higher sense of self-efficacy, coupled with better calibration of one's self-efficacy, enhanced predictiveness of performance, compared to higher self-efficacy alone. In other words, self-efficacy calibration "trimmed" away some inflated self-efficacy, which improved the predictiveness of math performance. As Schunk (1991) stated, high self-efficacy will not produce competent performance if students lack the

necessary skills. Thus, it was very possible that calibration adjusted for inflated self-efficacy and accounted for lack of necessary skills.

In the present study, the correlations between calibration accuracy and self-efficacy (r = -.01), and calibration bias and self-efficacy (r = .05) were insignificant. However, Pajares and Graham (1999) found otherwise; the correlations between accuracy and self-efficacy (fall, r = .67; spring, r = .48), and bias and self-efficacy (fall, r = .28; spring, r = .18) were statistically significant. However, Pajares and Graham (1999) used the same self-efficacy measure to compute accuracy and bias, which confounded these measures. When the present investigator used the same confounded methodology as did Pajares and Graham (1999), she found the correlation between calibration accuracy and self-efficacy to be significant (r = -.19, p = .05), as was that between calibration bias and self-efficacy (r = .28, p < .01). Thus, it appears that future research on calibration should not use the same measure of self-efficacy to avoid confounding.

In addition to having an effect on math performance, calibration accuracy also affected post-performance self-evaluation. Although calibration was hypothesized to have a direct effect on self-evaluation, it was found to have only an indirect effect mediated through self-efficacy; this negative indirect effect indicated that the higher the calibration accuracy, the lower the students' post-performance self-evaluation. This indicated an adjustment in students' perceived confidence after solving the problems. However, this negative indirect effect narrowly missed statistical significance at the .05 level (see Table 5). As for effect of calibration accuracy on effort judgment, it was directly negative, yet indirectly positive when mediated through self-efficacy beliefs. In

other words, the better-calibrated students exerted less effort, but those students with lower self-efficacy exerted more effort when solving math problems than high self-efficacy students. Since the total effect of calibration accuracy on effort judgments did not reach statistical significance (see Table 5), it was not a major contributing variable to effort judgment.

Hypothesis 2

The second hypothesis stated that self-efficacy and calibration accuracy had a positive influence on effort judgment. It was found, however, that self-efficacy had a negative direct effect on and negative correlation with effort judgment, thus disconfirming this hypothesis. This finding was contrary to prior self-efficacy research (Bong, 1997a; Pintrich et al., 1991, 1993; Schunk, 1983), indicating that perceived effort promoted higher self-efficacy and achievement. One possible explanation for this difference was that, in the present study, students reported self-perceptions of expended effort when they struggled with their performance rather than when they performed without difficulty. In supporting this interpretation, a linear effect of item difficulty on perceived effort was found in the present study.

The complex nature of post-performance effort judgment is also evident in prior research. Schunk (1983) measured perceived effort expenditure by asking students about their expended effort during their training sessions. Such a measure differed from that of the present study because the perceived effort was related to implementing training strategies (e.g., subtraction operations), whereas the present study measured students'

perceived effort for solving each math problem afterward. In Pintrich et al.'s (1991) Motivated Strategies for Learning Questionnaire (MSLQ), effort measures were also found to be positively correlated with academic self-efficacy (r = .44). However, the effort measure on the MSLQ was a set of four questions targeting students' global intended effort to complete study goals. Thus, it appears that performance context, such as judging effort intentionally before rather than after performance, greatly influenced the role of this variable.

Additional evidence indicated that effort judgments were influenced by other variables. For example, Brookhart and DeVoge (1999) found no relationship between self-efficacy and perceived mental effort. They examined the relationship between classroom assessments and student motivation and achievement. In their study, third graders were asked to self-report the amount of mental effort exerted with a number of language arts tasks (e.g., spelling test, vocabulary test, verb worksheet, language test, and so on) that eventually were used to assess their achievement and motivation. Students were asked about their perceived mental effort after completing each assessment task and before receiving teacher feedback. The effort questions, on a five-point Likert-type scale, consisted of the following: "How hard did you try..., How easy to understand..., How much did you concentrate..." (p. 412). Correlations between self-efficacy and perceived mental effort on the assessment tasks were mostly small and insignificant (Brookhart & DeVoge, 1999). Interviews were conducted with a number of subjects for each assessment task to further understand students' perceptions of the assessment. Based on the interviews, the researchers found that those students with extremely high self-

efficacy, who worked on unchallenging tasks, reported low perceptions of mental effort; the present study found similar results. It is possible that because their study did not quantitatively analyze task difficulty, the effect of self-efficacy and perceived effort was not shown.

Other evidence showed a negative relationship between self-efficacy and post-performance effort judgments. Zimmerman and Kitsantas (1999) found a negative relationship between post-performance effort attribution for negative performance outcome and self-efficacy in a study of high school students' writing/revision and self-regulatory processes. They found that students who attributed their rewriting deficiency to effort had lower self-efficacy, self-reactions, and intrinsic interest, whereas attribution to strategy use was positively related to these dependent measures. Thus, the present study's finding, that perceived effort was negatively correlated with self-efficacy, is not without historical precedent.

Although post-performance effort judgment was hypothesized to be influenced by self-efficacy and calibration accuracy, the present study showed that math performance had neither a direct nor an indirect effect on effort judgment. Prior research (Bong, 1997; Schunk, 1982, 1983; Vermeer et al., 2000) has shown that perceived effort and persistence (closely related to effort) have a positive relationship with performance. For example, using the MSLQ (Pintrich et al., 1991, 1993) to study high school student motivation and achievement, Bong (1997a) found that the effort measure correlated positively with students' reported course grades and academic self-efficacy. As mentioned earlier, the effort measure in the present study may have revealed the amount

of "struggle" or "frustration" students experienced on specific math problems rather than their willingness to expend effort beforehand. In addition, measuring effort has been shown to vary from global (Bong, 1997a; Pintrich et al., 1991; Schunk, 1993) to item-specific, as in the present study. In order to measure effort expenditure in future studies, the measure used needs to be refined and closely linked to testing situations and outcomes.

The main issue regarding the difference in findings between the present study and prior research (Bong, 1997a; Pintrich et al., 1991; Schunk, 1993) is timing—when the effort measure was assessed. Prior studies on effort generally reported a positive correlation with self-efficacy when effort is assessed before performance. These researchers assessed students' willingness to exert effort prior to engaging in performance. Such a predictive measure of effort produced a positive relationship with self-efficacy, which was also assessed prior to performance. However, in the present study, as well as in Zimmerman and Kitsantas (1999), effort judgment was measured after students had the opportunity to perform the math task. Students who had higher self-efficacy solved the problem more quickly, and thus were not required to exert all of their effort; this resulted in a negative correlation between self-efficacy and post-performance effort judgment.

Hypothesis 3

The third hypothesis investigated a possible gender difference on math self-efficacy, self-efficacy calibration, math performance, and post-performance self-

evaluation and effort judgment. The present study did not find support for this hypothesis from any of these measures. Although the present study found that, overall, boys rated themselves numerically higher in self-efficacy and post-performance self-evaluation, lower in effort judgment, and less calibrated than the girls (see Table 1), these differences in gender did not attain statistical significance. In addition, path analysis showed that the paths from gender to self-efficacy, self-evaluation, and effort judgment were not significant and were not included in the final path model. However, prior research on gender has been equivocal (Ewers & Wood, 1993; Pajares & Graham, 1999; Pajares & Kranzler, 1997).

With regard to calibration accuracy and calibration bias, both genders were equally overconfident or biased. Overconfidence was also found by Lundeberg, Fox, and Puncochar (1994) for both genders. However, they also found that males were significantly more overconfident than females, even when they had the wrong answers. By contrast, in the present study, although boys' self-efficacy was much higher than girls', this difference did not reach statistical significance. These findings concurred with prior studies conducted by Pajares and colleagues (Pajares & Graham, 1999; Pajares & Kranzler, 1995), showing that no gender differences were found regarding calibration accuracy or bias.

As for math performance, the present study did not find gender differences. Even though Hyde, Fennema, and Lamon (1990) found very slight gender differences in math performance, their meta-analysis showed that large differences between gender on math performance existed with older-age groups (high school and college-level), but not with

elementary or middle school groups. Ewers and Wood (1993) also found no gender difference with fifth graders on math performance. The present study concurred with the meta-analyses by Hyde, Fennema, and Lamon (1990) and Ewers and Woods (1993); no gender differences in math performance were found with this seventh-grade group. Although Pajares and Miller (1994) found gender differences in math performance, their study involved an older population, namely undergraduate college students.

Hypothesis 4

The fourth hypothesis stated that prior math achievement has a positive influence on math self-efficacy, self-efficacy calibration, and math performance. Previous math performance was based on students' sixth grade Iowa Test of Basic Skills (ITBS) scores. Path analysis revealed that students' previous math achievement had a significant direct effect ($\beta = .25$) on math performance, and indirect effects ($\beta = .40$) mediated via self-efficacy and calibration accuracy, roughly in equal magnitude. It appears that prior math skills primarily influenced self-efficacy beliefs and calibration of self-efficacy regarding subsequent math learning rather than math performance directly. Further, previous math achievement negatively correlated with self-efficacy calibration bias ($r = -.62$), suggesting that students with higher prior math achievement were less likely to be biased in their judgments. Similarly, students with higher previous math achievement were significantly higher on their self-efficacy calibration accuracy ($r = .63$).

Pajares and Graham (1999) revealed similar significant findings: students' previous achievement correlated positively with self-efficacy, calibration accuracy, and

math performance, and negatively with calibration bias. The present findings confirmed Bandura's (1986) social-cognitive hypothesis that students' previous mastery experiences provide them with information to gauge their perceived capability for future events. However, the present study went a step further to show that prior math achievement provided information for students not only to make future capability judgments (self-efficacy), but also to make more accurate and less biased ones.

Hypothesis 5

The fifth hypothesis stated that self-efficacy and self-efficacy calibration accuracy predicted post-performance self-evaluation. This hypothesis was partially confirmed (self-efficacy predicting post-performance self-evaluation was confirmed; calibration accuracy predicting post-performance self-evaluation was not confirmed). From the path analysis, self-efficacy was predictive of post-performance self-evaluation, with both direct and indirect effects mediated through effort judgment and math performance. Students who judged themselves capable of solving problems correctly were also likely to judge themselves the same way once they had the chance to solve the problems. Hacker et al. (2000) found similar results. Their study focused on undergraduate students' ability to predict and postdict a number of test performances in a classroom setting throughout an academic semester. The results showed undergraduate students' prior performance judgments had more weight in predicting their post-performance judgments than the actual performance itself. As shown in the present study, students'

self-efficacy beliefs had greater impact on their post-performance self-evaluation than the math performance itself (see Figure 1).

However, self-efficacy calibration accuracy did not have a direct effect on post-performance self-evaluation, as hypothesized; thus, the path was removed from the final path model. Interestingly, self-efficacy calibration accuracy had a negative indirect effect (which did not reach statistical significance) on post-performance self-evaluation mediated through self-efficacy. One possible explanation for this, based on the path analysis, was that better-calibrated students had slightly lower self-efficacy to begin with and, in turn, had slightly higher post-performance self-evaluation. On the other hand, students who had lower calibration accuracy rated their self-efficacy higher and, in turn, their post-performance self-evaluation lower. To test this hypothesis, the self-efficacy scores of high- and low-calibrated students were compared. With easy and moderately difficult items, both high- and low-calibration groups did not differ much on their self-efficacy scores (see Figure 1). However, as items became more difficult, better-calibrated students lowered their self-efficacy and became more sensitive to changes in difficulty level, but poorly-calibrated students were less likely to do so. Thus, the indirect effect of calibration accuracy on self-evaluation, which was mediated through self-efficacy, was negative.

Hypothesis 6

The sixth hypothesis stated that students have higher math self-efficacy and more accurate calibration on easier rather than more difficult problems. The findings of the

present study confirmed this hypothesis. Schraw et al. (1993) found that students had higher performance and more accurate judgments on easier tasks than on more difficult ones. However, their study measured calibration bias and accuracy using post-performance judgments. In addition, the easy items in Schraw et al.'s (1993) study included reading comprehension questions, whereas the difficult items included math solution questions. Thus, the comparison between easy and difficult items was, in fact, confounded between two different academic content areas rather than different items within one subject matter area. To avoid such measurement limitations, the present study tested students only with math problems ranging from easy to difficult. The multivariate repeated-measure analyses showed that students' ratings of their self-efficacy differed significantly according to levels of item difficulty. Item difficulty levels also affected students' calibration (accuracy and bias), effort judgment, and self-evaluation. Students were more accurate and less biased in their self-efficacy on easier math items than on moderately difficult or difficult items.

There were also calibration group and item difficulty level interactions. In terms of post-performance self-evaluation, the high-calibrated group was slightly lower than the low-calibrated group when the items were easy. However, with moderately difficult items, both groups' self-evaluation scores dropped, with the low-calibrated group dropping more than the high-calibrated group. With difficult items, both groups' self-evaluation scores dropped even more, but did not show much difference (see Figure 2). Since the self-evaluation was measured after math performance, it was possible that low-calibrated students' self-evaluation scores reflected their poorer math performance. To

test this hypothesis, both calibrated groups' math performances were calculated according to item difficulty level (see Figure 3). Figure 3 showed that the difference between the two groups was most noticeable on moderately difficult items, which might explain why the low-calibrated group's post-performance self-evaluation scores dropped more than the high-calibrated group's on moderately difficult items. It is important to note that even the low-calibrated group's self-evaluation scores decreased, reflecting a reaction to item difficulty levels.

Summary

This study advanced an understanding of how students perceive their capabilities according to task difficulty level and judge their effort in solving math problems. First, it was important to establish self-efficacy calibration as a separate dimension of self-efficacy. Calibration measures improved the predictiveness of self-efficacy beliefs. Students who expressed a higher sense of capability performed better (Bandura, 1986); however, the present study showed that more accurate perceptions of capability were associated with improved math performance. Second, this study revealed that students who had a high sense of self-efficacy actually performed better. However, students who were poorly calibrated overestimated their self-efficacy. Although higher self-efficacy is desirable, as the self-efficacy research has well documented (Bandura, 1986; Schunk, 1991), accurately calibrated beliefs are also important to promote in students. By investigating students' self-efficacy and calibration in terms of task difficulty level, the

present study identified that students were more sensitive to their performance on easier tasks than on more difficult ones.

Third, boys and girls did not differ, overall, on measures such as math performance, self-efficacy judgments, post-performance self-evaluation and effort judgement, and calibration accuracy and bias. Although numerical differences were found based on gender, they were not statistically significant.

Fourth, this study revealed important information on the effort judgments made by students while solving math problems. Such task-specific information revealed how students judge their effort across particular problems. Unfortunately, in the present study, effort judgment proved to be a measure of the "amount of struggle" rather than a measure of intended effort on each math problem. The latter meaning was attributable to the fact that the present study was a first attempt to measure students' perceived effort on an item-per-item basis.

Educational Implications

This study indicated that higher self-efficacy was desirable, but calibration accuracy combined with self-efficacy was ideal. This research does not mean to minimize the importance of students' confidence in their capability; rather, it poses the important question that when confidence does not correspond with high capability, how can educators improve students' self-efficacy calibration? Zimmerman, Bonner, and Kovach (1996) provided a cyclical model to guide educators in assisting their students to become more self-regulated learners. To improve students' self-efficacy judgments of

their academic skills, these researchers asked them to estimate their efficacy before undertaking brief class quizzes. Students then graphed their self-efficacy judgments along with the actual scores they obtained. Once noted, the disparity readily guided students to adjust their subsequent self-efficacy judgments to attain greater accuracy. An educational implication here is to use self-recording to help students better assess their math skills and to adjust studying appropriately.

Another educational implication is the importance of measurement issues regarding research on concepts of self-perception. The results of the present study regarding the negative correlation between self-efficacy beliefs and effort judgment were not consistent with prior research (Bong, 1997; Pintrich et al., 1991; Schunk, 1982, 1983) which reported a positive correlation. An attempt was made to resolve this discrepancy by reviewing differences in how effort was measured; the measurement issue seemed to be the most reasonable one to explain the difference in findings. The implication here for educational researchers is that methods for measuring and interpreting effort judgments as well as self-efficacy beliefs need careful attention.

Directions for Future Research

One future research topic would be to examine how students develop their self-efficacy calibrations. A cross-sectional developmental study could shed light on how students' self-efficacy calibrations vary by age of student. Another future research topic could compare students' self-efficacy calibration to their learning strategies. Are students aware of the learning strategies they use? Students who overestimate their strategy use

might also overestimate their self-efficacy levels. For future research on the motivational measure of effort in relation to self-efficacy, one should consider how effort is measured. When students were assessed on their perceived effort expenditure, timing greatly influenced the results and interpretations, as demonstrated in the present study. Finally, for future studies, one could develop and assess the effectiveness of training protocols designed to assist students in developing highly-calibrated self-efficacy.

Conclusion

This study provided a detailed assessment of a range of self-regulatory processes or beliefs that was hypothesized to underlies students' sense of agency about acquiring math competence. It explored the critical relationship between variations in context (e.g., specific math problems and their difficulty level) and students' personal processes (e.g., judgments of their math capabilities, self-efficacy calibration, effort judgment, and performance self-evaluation). The results revealed clearly that students' self-efficacy beliefs play a pivotal role in their acquisition of mathematical competence. Such information can be vital in assisting educators to tailor interventions that will enhance students' beliefs on their capability to learn math and as well as the actual success.

Appendix A

Form 1 of Student Packet

Form 1: _____

ANSWER SHEET

Please circle your response by choosing a number

Example 1

How confident are you about solving this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

Example 2

How confident are you about solving this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

---

1. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

2. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

3. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

4. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

5. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

6. ***How confident are you about solving this math problem correctly?***

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

7. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

8. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

9. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

10. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

11. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

12. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

13. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

14. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

15. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

Form 1

*Please **solve** each math problem in the space provided. When you finish a problem, please write the **answer** on the blank line indicated and **circle** your **responses** for the 2 questions on the bottom of each page.*

Example 1

What is 35 plus 5.5?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

Form 1: _____

Please **solve** each math problem in the space provided. When you finish a problem, please write the **answer** on the blank line indicated and **circle** your **responses** for the 2 questions on the bottom of each page.

1.

A newspaper reported that about 18,189 trees had been planted in the park. What is the number of trees rounded to the nearest hundred?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

2.

In a discus-throwing competition, the winning throw was 61.60 m. The second-place throw was 59.72 m. How much longer was the winning throw than the second-place throw?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

3.

Two groups of tourists each have 60 people. If 3/4 of the first group and 2/3 of the second group board buses to travel to a museum, how many more people in the first group board buses than in the second group?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

4.

Anna had a bag of marbles. She gave half of them to James and then a third of the marbles still in the bag to Pat. She then had 6 marbles left. How many marbles were in the bag to start with?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

5.

A class has 28 students. The ratio of girls to boys is 4 : 3. How many girls are in the class?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

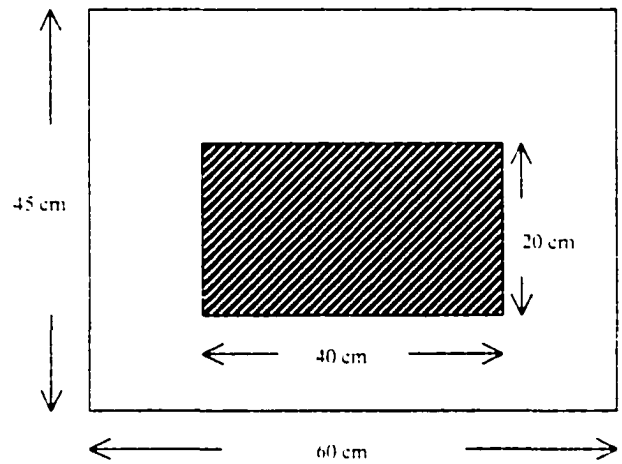
6.

In a quadrilateral, two of the angles each have a measure of 110°, and the measure of a third angle is 90°. What is the measure of the remaining angle?

Answer: _____

A. How much effort did you put in to solve this problem?
1 **2** **3** **4** **5** **6** **7** **8** (none) (a little) (a lot) (all)
B. How confident are you that you solved this math problem correctly?
1 **2** **3** **4** **5** **6** **7** **8** (not at all) (somewhat) (mostly) (completely)

7.



A shaded rectangular picture is pasted on a sheet of white paper as shown above. What is the area of the white paper not covered by the picture?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

8.

What does 6000 minus 2369.4 equal? Please show your work.

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

9.

A group of students has a total of 29 pencils and everyone has at least one pencil. Six students have 1 pencil each, 5 students have 3, and the rest have 2. How many students have only 2 pencils?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

10.

If the price of a can of beans is raised from 60 to 75 cents, what is the percent increase in the price?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

11.

Janis, Maria, and their mother were eating a cake. Janis ate 1/4 of the cake. Maria ate 1/3 of the cake. Their mother ate 1/3 of the cake. How much of the cake is left?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

12.

The numbers in the sequence 2, 7, 12, 17, 22, ... increase by fives. The numbers in the sequence 3, 10, 17, 24, 31, ... increase by seven. The number 17 occurs in both sequences. If the two sequences are continued, what is the next number that will occur in both sequences?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

13.

Four children measured the width of a room by counting the number of paces it took them to cross it. The chart shows their measurements. Who had the longest pace?

Name	Number of Paces
Stephen	10
Arlene	7
Ana	9
Carlos	8

Answer: _____

A. How much effort did you put in to solve this problem?

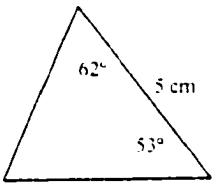
1 (none) **2** **3** (a little) **4** **5** (a lot) **6** **7** **8** (all)

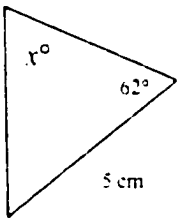
B. How confident are you that you solved this math problem correctly?

1 (not at all) **2** **3** (somewhat) **4** **5** (mostly) **6** **7** **8** (completely)

14.

These triangles are congruent. The measures of some of the sides and angles of the triangles are shown. What is the value of x ?





Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

15.

A drawer contains 28 pens: some white, some blue, some red, and some gray. If the probability of selecting a blue pen is 2/7, how many blue pens are in the drawer?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

Appendix B

Form 2 of Student Packet

Form 2: _____

ANSWER SHEET

Please circle your response by choosing a number.

Example 1

How confident are you about solving this math problem correctly?

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

Example 2

How confident are you about solving this math problem correctly?

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

---

1. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

2. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

3. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

4. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

5. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

6. ***How confident are you about solving this math problem correctly?***

1 (not at all) 2 3 (somewhat) 4 5 (mostly) 6 7 (completely) 8

7. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

8. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

9. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

10. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

11. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

12. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

13. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

14. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

15. ***How confident are you about solving this math problem correctly?***

1 2 3 4 5 6 7 8
(not at all) (somewhat) (mostly) (completely)

Form 2

*Please **solve** each math problem in the space provided. When you finish a problem, please write the **answer** on the blank line indicated and **circle** your **responses** for the 2 questions on the bottom of each page.*

Example 1

What is 44 plus 7.3?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

B. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

Form 2: _____

Please **solve** each math problem in the space provided. When you finish a problem, please write the **answer** on the blank line indicated and **circle** your **responses** for the 2 questions on the bottom of each page.

1.

A newspaper reported that about 19,179 trees had been planted in the park. What is the number of trees rounded to the nearest hundred?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

2.

In a discus-throwing competition, the winning throw was 62.50 m. The second-place throw was 59.63 m. How much longer was the winning throw than the second-place throw?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

3.

Two groups of tourists each have 60 people. If 4/5 of the first group and 2/3 of the second group board buses to travel to a museum, how many more people in the first group board buses than in the second group?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

4.

Anna had a bag of marbles. She gave half of them to James and then a quarter of the marbles still in the bag to Pat. She then had 6 marbles left. How many marbles were in the bag to start with?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

5.

A class has 28 students. The ratio of girls to boys is 5 : 2. How many girls are in the class?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

6.

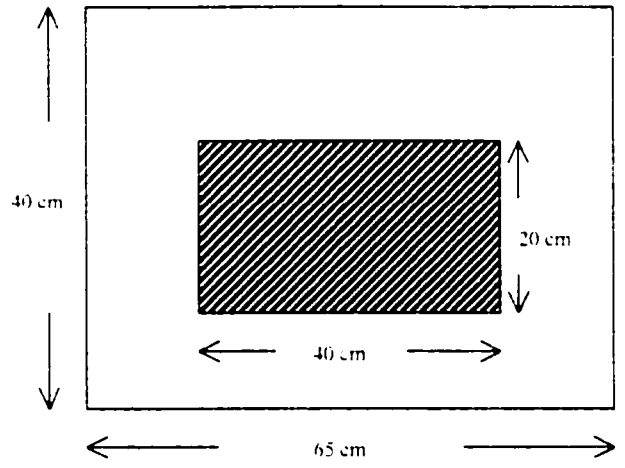
In a quadrilateral, two of the angles each have a measure of 110°, and the measure of a third angle is 80°. What is the measure of the remaining angle?

Answer: _____

A. How much effort did you put in to solve this problem?							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(not at all)		(somewhat)		(mostly)		(completely)	

7.



A shaded rectangular picture is pasted on a sheet of white paper as shown above. What is the area of the white paper not covered by the picture?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

8.

What does 7000 minus 2479.6 equal? Please show your work.

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

9.

A group of students has a total of 29 pencils and everyone has at least one pencil. Eight students have 1 pencil each, 5 students have 3, and the rest have 2. How many students have only 2 pencils?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

10.

If the price of a can of beans is raised from 60 to 72 cents, what is the percent increase in the price?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

11.

Janis, Maria, and their mother were eating a cake. Janis ate 1/3 of the cake. Maria ate ¼ of the cake. Their mother ate ¼ of the cake. How much of the cake is left?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

12.

The numbers in the sequence 6, 11, 16, 21, 26, ... increase by fives. The numbers in the sequence 3, 9, 15, 21, 27, ... increase by six. The number 21 occurs in both sequences. If the two sequences are continued, what is the next number that will occur in both sequences?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

13.

Four children measured the width of a room by counting the number of paces it took them to cross it. The chart shows their measurements. Who had the longest pace?

Name	Number of Paces
Stephen	12
Arlene	9
Ana	7
Carlos	8

Answer: _____

A. How much effort did you put in to solve this problem?

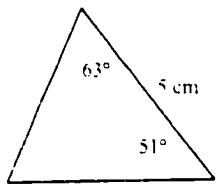
1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

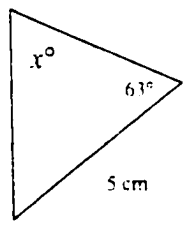
C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

14.

These triangles are congruent. The measures of some of the sides and angles of the triangles are shown. What is the value of x ?





Answer: _____

A. How much effort did you put in to solve this problem?							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
(not at all)		(somewhat)		(mostly)		(completely)	

15.

A drawer contains 28 pens: some white, some blue, some red, and some gray. If the probability of selecting a blue pen is 4/7, how many blue pens are in the drawer?

Answer: _____

A. How much effort did you put in to solve this problem?

1	2	3	4	5	6	7	8
(none)		(a little)		(a lot)			(all)

C. How confident are you that you solved this math problem correctly?

1	2	3	4	5	6	7	8
(not at all)		(somewhat)		(mostly)		(completely)	

REFERENCES

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

Bandura, A. (1997). Self-efficacy: The exercise of control. New York: W. H. Freeman and Company.

Bong, M. (1997a). Congruence of measurement specificity on relations between academic self-efficacy, effort, and achievement indexes. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Bong, M. (1997b). Generality of academic self-efficacy judgments: Evidence of hierarchical relations. Journal of Educational Psychology, 89, 696-709.

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

Brookhart, S. M., & DeVoge, J. G. (1999). Testing a theory about the role of classroom assessment in student motivation and achievement. Applied Measurement in Education, 12, 409-425.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. Psychology of Women Quarterly, 18, 585-609.

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, 259-267.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology (3rd ed.). Boston: Allyn & Bacon.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction (6th ed.). New York: Longman.

George, D., & Mallery, P. (2001). SPSS for windows step by step: A simple guide and reference (3rd ed.). Boston: Allyn & Bacon.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test predication and performance in a classroom context. Journal of Educational Psychology, 92, 160-170.

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

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. Psychology of Women Quarterly, 14, 299-324.

Impara, J. C., & Plake, B. S. (Eds.). (1998). Mental measurement yearbook (13th ed.). Lincoln, NB: Buros Institute of Mental Measurements.

Joreskog, K., & Sorbom, D. (1996). LISREL 8: User's reference guide. Chicago: Scientific Software International, INC.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. Journal of Educational Psychology, 86, 114-121.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. Journal of Counseling Psychology, 38, 30-38.

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

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). Self-efficacy beliefs and general mental ability in mathematical problem-solving. Contemporary Educational Psychology, 20, 426-443.

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, 193-203.

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. The Journal of Experimental Education, 65, 213-228.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ) (Tech. Rep. No. 91-B-004). Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). Educational and Psychological Measurement, 53, 801-813.

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

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

Schunk, D. H. (1982). Effects of effort attributional feedback on children's perceived self-efficacy and achievement. Journal of Educational Psychology, 74, 548-556.

Schunk, D. H. (1983). Ability versus effort attributional feedback: Differential effects on self-efficacy and achievement. Journal of Educational Psychology, 75, 848-856.

Schunk, D. H. (1991). Self-efficacy and academic motivation. Educational Psychologist, 26, 207-231.

TIMSS (1995). Mathematics items: Released set for population 2 (seventh and eighth grades). IEA Third International Mathematics and Science Study (TIMSS). Chestnut Hill, MA: Boston College.

Vermeer, H. J., Boekaerts, M., & Seegers, G. (2000). Motivational and gender differences: Sixth-grade students' mathematical problem-solving behavior. Journal of Educational Psychology, 92, 308-315.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. Contemporary Educational Psychology, 25, 82-91.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). Developing self-regulated learners: Beyond achievement to self-efficacy. Washington, DC: American Psychological Association.

Zimmerman, B. J., & Kitsantas, A. (1999). Acquiring writing revision skill: Shifting from process to outcome self-regulatory goals. Journal of Educational Psychology, 91, 241-250.