

Summer 2012

## Can Practice Calibrating by Test Topic Improve Public School Students' Calibration Accuracy and Performance on Tests?

Rose M. Riggs  
*Old Dominion University*

Follow this and additional works at: [https://digitalcommons.odu.edu/teachinglearning\\_etds](https://digitalcommons.odu.edu/teachinglearning_etds)



Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Educational Psychology Commons](#)

---

### Recommended Citation

Riggs, Rose M.. "Can Practice Calibrating by Test Topic Improve Public School Students' Calibration Accuracy and Performance on Tests?" (2012). Doctor of Philosophy (PhD), Dissertation, Teaching & Learning, Old Dominion University, DOI: 10.25777/pj9w-th02  
[https://digitalcommons.odu.edu/teachinglearning\\_etds/38](https://digitalcommons.odu.edu/teachinglearning_etds/38)

This Dissertation is brought to you for free and open access by the Teaching & Learning at ODU Digital Commons. It has been accepted for inclusion in Teaching & Learning Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact [digitalcommons@odu.edu](mailto:digitalcommons@odu.edu).

CAN PRACTICE CALIBRATING BY TEST TOPIC IMPROVE PUBLIC SCHOOL  
STUDENTS' CALIBRATION ACCURACY AND PERFORMANCE ON TESTS?

by

Rose M. Riggs  
B.S. May 1988, Syracuse University  
M.S. August 2004, Old Dominion University

A Dissertation Submitted to the Graduate Faculty of  
Old Dominion University in Partial Fulfillment of the  
Requirements for the Degree of

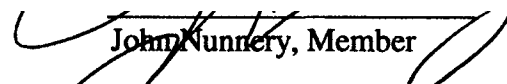
DOCTOR OF PHILOSOPHY

CURRICULUM AND INSTRUCTION

OLD DOMINION UNIVERSITY  
August 2012

Approved by:

✓   
Linda Bol, Director

  
John Nunnery, Member

  
Daniel Dickerson, Member

## ABSTRACT

### CAN PRACTICE CALIBRATING BY TEST TOPIC IMPROVE PUBLIC SCHOOL STUDENTS' CALIBRATION ACCURACY AND PERFORMANCE ON TESTS?

Rose M. Riggs  
Old Dominion University, 2012  
Director: Dr. Linda Bol

The effect of a calibration strategy requiring students to predict their scores for each topic on a high stakes test was investigated. The utility of self-efficacy towards predicting achievement and calibration accuracy was also explored. One hundred and ten sixth grade math students enrolled in an urban middle school participated. Students were assigned to either a calibration practice group or a no practice condition. Students in the practice condition completed a self-efficacy scale specific to math at the beginning of the study. They also practiced making predictions for each topic on each of three tests over a three month period to determine if their calibration accuracy and performance on tests would be increased. Students in both the practice and no-practice conditions calibrated their scores topically on the final, high stakes math test at the end of the course. There was not a significant difference between the conditions in calibration accuracy on the final, high stakes test, indicating that calibration practice did not improve accuracy. There was no significant difference between the practice and no practice conditions in on achievement. However, a significant relationship was found between achievement level and calibration accuracy. Higher achieving students in both the calibration practice and

no practice conditions were significantly more accurate than lower achieving students in both conditions. Self-efficacy was not found to be predictive of achievement or calibration accuracy. Further research is needed to identify more effective strategies for enhancing metacognitive judgments, self-efficacy, and performance.

## ACKNOWLEDGEMENTS

I am very grateful to my committee for their help and suggestions with my dissertation. However, a special thanks goes to the person that helped me the most, my committee chair Dr. Linda Bol, whose constant support, encouragement, and editing were above and beyond what anyone could expect. I am so glad she agreed to be my chair. I would also like to thank my family who had to endure the many years of labor and often frustration that this took for me to complete. Thank you to my daughter Annie who inspires me with her scholarly dedication and ages-old wisdom, to Jacen, my adventurous and thoughtful son, and to Laura, my daughter who keeps me young. Finally, a very special thank you to John, who served as my technical advisor and coach, in addition to his long-standing role of life partner and friend.

## TABLE OF CONTENTS

LIST OF TABLES.....	vii
LIST OF FIGURES.....	viii
Chapter	
I. INTRODUCTION.....	1
WHY IMPROVING STUDENTS' MATH PERFORMANCE IS CRITICAL .....	3
THE SIGNIFICANCE OF HIGH STAKES TESTING .....	5
THE RELATIONSHIP BETWEEN SELF-EFFICACY AND CALIBRATION ....	11
SIGNIFICANCE AND PURPOSE OF STUDY .....	16
OVERVIEW OF METHOD .....	19
II. LITERATURE REVIEW.....	20
OVERVIEW .....	20
CALIBRATION TYPES AND MEASURES .....	20
THE RELATIONSHIP BETWEEN CALIBRATION AND SELF-EFFICACY ....	31
THE RELATIONSHIP BETWEEN CALIBRATION ACCURACY, ACHIEVEMENT LEVEL, AND BIAS .....	35
CALIBRATION STRATEGIES .....	40
RATIONALE FOR STUDY.....	48
RESEARCH QUESTIONS AND HYPOTHESES .....	50
SUMMARY AND CONTRIBUTIONS OF PRESENT STUDY .....	52
III. METHODOLOGY .....	54
INTRODUCTION .....	54
SCHOOLS AND PARTICIPANTS .....	55
DESIGN.....	56
MEASURES .....	57
PROCEDURE.....	60
IV. RESULTS .....	64
INTRODUCTION .....	64
IMPACT OF CALIBRATION PRACTICE ON ACHIEVEMENT .....	65
ACHIEVEMENT BY TREATMENT CONDITION.....	68

IMPACT OF CALIBRATION PRACTICE ON CALIBRATION ACCURACY...	70
CALIBRATION ACCURACY BY TREATMENT CONDITION .....	72
IMPACT OF ACHIEVEMENT LEVEL ON CALIBRATION ACCURACY.....	73
RELATIONSHIP BETWEEN SELF-EFFICACY, ACHIEVEMENT, AND CALIBRATION ACCURACY .....	80
STUDENT PERCEPTIONS OF TREATMENT EFFECTIVENESS.....	82
V. DISCUSSION .....	85
OVERVIEW .....	85
CALIBRATION PRACTICE AND TEST PERFORMANCE .....	85
CALIBRATION PRACTICE AND CALIBRATION ACCURACY .....	89
ACHIEVEMENT LEVEL AND CALIBRATION ACCURACY .....	92
SELF-EFFICACY AND PERFORMANCE .....	94
SELF-EFFICACY AND CALIBRATION ACCURACY .....	96
STUDENT PERCEPTIONS OF CALIBRATION.....	98
LIMITATIONS.....	100
FUTURE DIRECTIONS FOR RESEARCH.....	101
EDUCATIONAL IMPLICATIONS.....	103
SUMMARY AND CONCLUSIONS .....	104
REFERENCES .....	106
<u>APPENDICES</u>	
APPENDIX A: TEST CALIBRATION BY TOPIC (PRACTICE CONDITION) .....	11716
APPENDIX B: STUDENT OPT-OUT FORM .....	117
APPENDIX C: PATTERNS OF ADAPTIVE LEARNING SCALES (PALS)....	118
APPENDIX D: INSTRUCTIONS FOR EFFICACY SCALE PROCTORING ....	119
APPENDIX E: TEACHER INSTRUCTIONS TO STUDENTS .....	120
APPENDIX F: FINAL SOL TEST CALIBRATION BY TOPIC .....	121
APPENDIX G: STUDENT PERCEPTIONS OF STRATEGY QUESTIONS.....	122
VITA.....	123

## LIST OF TABLES

Table	Page
1. Average SOL Scores for Calibration Practice and No Practice Groups.....	67
2. ANOVA and MANOVA Results for Calibration Practice and No Practice Groups on SOL Test .....	69
3. Descriptive Statistics for Calibration Accuracy and Direction between Calibration Practice and No Practice Groups on SOL Test .....	71
4. ANOVA and MANOVA Results for SOL Calibration Accuracy.....	73
5. Effect of Achievement (high or low) and Practice Condition on SOL Total Calibration Accuracy .....	74
6. Total Calibration Accuracy Means and Standard Deviations for High/Low Achievement Groups .....	76
7. Multiple Analyses Of Variance showing Effect of Achievement Level (high or low) and Practice Condition on SOL Topical Calibration Accuracy.....	78
8. Topical Calibration Accuracy Descriptive Statistics: High and Low Achievement Groups.....	79
9. Descriptive Statistics for Self Efficacy Scale Question Answers.....	81
10. Regression: Self-Efficacy Belief as a Predictor of Achievement or Calibration Accuracy.....	82
11. Descriptive Statistics for Student Perceptions of Treatment Effectiveness Answers.....	84



## LIST OF FIGURES

Figure	Page
1. Absolute Calibration Accuracy by Achievement Level and Practice Condition..	77

Can Practice Calibrating by Test Topic Improve Public School Students' Calibration  
Accuracy and Performance on Tests?

CHAPTER I

INTRODUCTION

Benjamin Spock said “Trust yourself. You know more than you think you do” (p.1). Unfortunately, while this may or may not be true for parenting, the concept does not seem transferrable to most students’ judgments of their own learning. It often seems that students trust themselves a bit too much when judging their knowledge of a subject. Some students are better than others at judging their own knowledge. These learners are academically independent, and have most likely learned critical self-evaluation skills. Dembo and Eaton (2000) suggest that students learn to be academically independent when they learn how to regulate their own behaviors to control the outcome of their performance. Successful learners monitor and control their behaviors when given a learning task. Unfortunately, some students lack awareness of their own learning and are not as likely to monitor and control their behaviors. When they do try to monitor their own learning, they often have faulty self-evaluation skills. Faulty self-evaluation skills may result in a mismatch between how much students think they know and how much they actually know. This may result in poorer performance outcomes than the student may have expected. Self-regulation influences students’ ability to judge how well they will perform on a task like an exam, and the accuracy of this judgment is termed

calibration. Bol & Hacker (2001) explain calibration as a students' ability to judge how well they will perform before taking a test (prediction) and then how well they performed after completing the test (postdiction). Calibration is measured in different ways as discussed later. The intent of this research was to measure the impact of a calibration strategy on middle school students' calibration accuracy and test performance, and determine if self-efficacy was a significant predictor of calibration accuracy or test performance.

Middle school students are at a transitional period in their lives. Adolescents are learning more advanced metacognitive skills, such as self-regulation, at the same time they are developing their self-efficacy beliefs in academic subjects, and the two are often connected. For example, if a student believes they are incapable of performing well in a subject (e.g. "I'm just bad at math") they may be less likely to consider ways to improve their performance in that subject. Helping these students develop the ability to self-regulate through accurate self-evaluation is important in a climate of high-stakes testing and particularly in math (Lynn, 2008).

The related research that provides the rationale for the present study includes the importance of enhancing students' math performance, the significance of high stakes testing such as Virginia's Standards of Learning (SOL) Tests, and adolescent self-efficacy, and calibration as it relates to math and testing. An overview of the relevant literature is presented in this chapter. A more in-depth review that includes descriptions of related studies is provided in Chapter II.

### Why Improving Students' Math Performance is Critical

Math has always been one of the core subjects tested, and it is becoming increasingly important as our society grows more technically sophisticated. It is one of the focal points of STEM (Science, Technology, Engineering, and Math) initiatives, and it is a prominently tested subject for high stakes tests (Pajares & Graham, 1999). Many fear that America is losing its competitive edge as our students' scores in math continue to lag behind other countries.

This fear is not without basis according to the report on Comparative Indicators of Education in the United States and Other G-8 Countries: 2009. Group of Eight (G-8) countries include Canada, France, Germany, Italy, Japan, the Russian Federation, the United Kingdom, because these are the most economically developed countries and among the United States' largest economic partners (p.iii). This report compiled statistics from the 2007 Trends in International Mathematics and Science Study (TIMSS). The TIMSS indicates that American students' scores are consistently lower than both Japan's and The Russian Federation's students, and only average in comparison with other countries (Miller, Anindita, Malley & Burns, 2009). In addition, according to a report compiled in 2000 from the National Center for Education Statistics, 35% of freshmen entering public 2-year and 16% of freshmen entering public 4-year degree-granting institutions required remediation in math, representing an average of 26%. The number of students requiring reading and writing remediation was less, at 13% and 16% respectively. This indicates that many of our students are not getting a solid mathematical foundation.

The NAEP Trends Report published in 2009 reveals the mathematical concepts that high school students have difficulty with, such as percents, finding area, estimations, and simple algebra, are all concepts taught in most middle school math curriculums. The report also indicated that although achievement gaps are generally shrinking, in some instances they are growing. In high school math, the gap actually increased between the percentage of low-income and non-low-income students who obtained a proficient rating on tests.

The same report showed that gaps were largest in high school math and smallest in elementary school math, and that the largest gaps were between the African American versus White subgroups: “In high school math, for example, the mean (average) percentage proficient was 45% for the African American subgroup and 74% for the white subgroup, resulting in a black-white gap of 29 percentage points” (p.17). These gaps are troubling considering that a common goal of many school divisions across the nation is reducing the achievement gap between racial groups.

Haberman, a leader in urban education, contends that the achievement gap has widened since 1962. He ranks the U.S. White students’ achievement in math as 7th worldwide, while Black and Hispanic students’ achievement ranks 27th worldwide. Strategies to improve students’ math performance are clearly necessary, particularly in the 120 largest urban school districts. These districts educate 11 million students who are mostly minorities or of poverty, and at the bottom end of the achievement gap (2004).

If the United States is to remain globally competitive in an increasingly technological society, it is necessary to have a workforce that has critical math and

science skills. Beyond ensuring the country's competitiveness, STEM subjects are important for students who want "a decent wage-paying job in the economy of the 21st century" (Morrison & Bartlett, 2009)

### The Significance of High Stakes Testing

High stakes testing is quite controversial despite its widespread usage. Advocates of high stakes tests argue that the standards provide incentives for students and increase the value of a high school diploma, whereas opponents suggest they cause too much stress and may lead to students dropping out of school altogether (Papay, Murnane, & Willett, 2008). Hacker, Bol, and Keener (2009) assert that student performance on these tests has an impact on educational placements, grade promotion, graduation, college admissions, and eventually into entry of various professions (p. 438).

Some suggest that The No Child Left Behind (NCLB) Act pushes low performers, who are often minorities and disadvantaged students, out of schools to bolster test scores. States are allowed to determine how to compute their own drop-out rates. This results in misleading numbers due to variances in reporting methods across states. Many students are not counted as drop-outs because they legitimately cannot be found after leaving the school. In addition, schools are allowed to set their own goals relating to graduation rates, so any rate of progress is acceptable (Swanson, 2004).

Papay et al. (2008) found that exit examinations, such as those required for high school math, prevent students from graduating from high school because fear of failing may cause them to drop out before even taking the test, or if they fail the examination, they may drop out before re-taking it. Students who do retake the test and still fail it

multiple times cannot graduate. Low-income urban students who fail the exit mathematics examination are just as likely as suburban students to retake the test, but they are much less likely to pass on retest, resulting in their inability to graduate.

It is necessary to have an understanding of The No Child Left Behind (NCLB) Act to appreciate the extent of the pressure that high-stakes tests cause. NCLB stresses accountability at all levels and standards for student achievement. To obtain federal funding, states are required to develop assessments in basic skills that are administered to all students in certain grades. Title 1 provides federal funds to schools serving at-risk children in high poverty areas. These schools rely on assessment results to show progress among these students and avoid punitive consequences. Consequences include having the state take over the school or close the school completely, in addition to the public embarrassment schools face when their rankings are publicized.

NCLB requires:

- Annual testing in grades 3-8 and at least once in high school to measure student progress in reading and mathematics.
- Schools, school divisions and states to meet annual objectives for Adequate Yearly Progress (AYP) for student performance on statewide tests in reading and mathematics.
- The identification of states, schools and school divisions making and not making AYP.
- All students to be proficient in reading and mathematics by 2013-2014 (United States Department of Education, 2010).

As a result, these high-stakes tests are often key factors in determining grade promotion or retention and graduation from high school, in evaluating schools on the basis of students' scores, and in the individual evaluation of teachers based on student scores (Popham, 2001; Ryan, Ryan, & Arbuthnot, 2007). Despite lingering controversies over their usefulness for the past century (Linn, 2001) and their ability to improve education or reduce the achievement gap (Mathis, 2003), most states, including Virginia, have mandated standardized, criterion-referenced tests for core subjects, including math, in grades 3-8 and high school. Student performance on these tests impacts everyone from the State Departments of Education to the individual student.

Virginia has developed an accountability system, based in large part on state-developed standards aligned with their assessments. According to the Accountability Guide published on Virginia's Department of Education website, Virginia has implemented an accountability system based on rigorous academic standards, known as the Standards of Learning (SOL), and progress on these standards is measured through annual assessments of student achievement (2010). Also according to the Virginia Department of Education, schools must make Adequate Yearly Progress (AYP) by meeting increasingly higher objectives for pass rates on the SOL tests.

Middle and high schools receive one of four accreditation ratings each year based on the achievement of students on SOL assessments and other tests in English, history, mathematics and science taken during the previous academic year. The accreditation ratings are: fully accredited, accredited with warning, accreditation denied, and conditionally accredited (for new schools). Accreditation denied is the result of four



years of failure to meet full accreditation. These schools undergo increasingly stringent sanctions each year they fail to be accredited. If a Title 1 school is denied accreditation for four consecutive years, they must choose whether to reopen the school as a charter school, replace all or most of the school staff relevant to the school's failure to make AYP, turn the management of the school over to a private educational management company, or arrange other major restructuring of school governance (Virginia Department of Education, 2011).

Not surprisingly, anxiety and stress are common amongst teachers nationwide who feel pressured to ensure their students pass these high stakes tests (Barksdale-Ladd & Thomas, 2000; Bol, 2004). As a result, teachers often focus on preparing their students for the yearly SOL tests, resulting in drill and practice with tests similar to the SOLs. Bol's 2004 study on teacher's assessment practices in high-stakes environments confirmed this finding. Bol surveyed classroom teachers on the influence of the SOLs on assessment and teaching practices. Responses indicated that the teachers align their assessments and instruction with the SOLs. They spent class time working on SOL practice problems, and they were concerned about how their students performed on the SOLs. And teachers are not the only ones under stress. The pressure to pass these tests places an additional burden on groups that are already often struggling in school, such as low-income and special needs students (Papay et al, 2008).

This leads to another concern many have with high-stakes testing. They do not encourage critical thinking and metacognitive skills. Bol and Nunnery (2004) argue that whole school reforms that bolster learning for at-risk students have been ineffective

because of competing priorities. Teachers reported that they do not use authentic performance and alternative assessments as readily as they may have in the past due to the to teach to multiple-choice high stakes tests (Bol & Nunnery). Rote drill and practice is a common technique for preparing students for the tests, but this does not help students develop critical thinking skills. As a result, strategies that improve metacognition and foster critical thinking are especially needed now for at risk students, who are probably “the real losers” in high-stakes testing environments.

The Norfolk Public School District in Southeastern Virginia serves an urban population of disadvantaged and minority children. It is subject to the NCLB requirements for funding, including high-stakes testing (Virginia Department of Education Website: Accreditation & AYP, 2011). Information from the State’s website indicates that student demographics are 63% African American, 23% Caucasian, 6% Hispanic, 6% self-identified from two races, and 2-3% Asian or Other. Over half—56% of students enrolled in NPS, are eligible for free lunch, with another 8% eligible for reduced price lunches. (Virginia Department of Education Website: Enrollment & Demographics, 2011). Norfolk Public School students graduating with standard diplomas must pass one of three math End-of-Course (EOC) SOL tests in high school; either Algebra, Algebra II, or Geometry, while students graduating with advanced diplomas must pass at least two out of three of those math EOC SOL tests.

Data from the Virginia Department of Education’s website indicates that during the 2008-2009 school year, 10% of students that took the Algebra 1 SOL exam failed, 24% of students that took the Geometry SOL failed, and 12% of students that took the

Algebra II SOL exam failed. All three of these failure rate percentages are above the average for the State's fail rates. In addition, these rates do not necessarily reflect the number of students who passed these tests the first time they took them; students may retake these tests during the school year, and the achievement of students on all retakes of end-of-course assessments in reading and mathematics are included in the calculation of AYP ratings. If a student fails a test required for graduation and successfully retests during the same school year, the first test does not count in calculating AYP.

The Norfolk School District still has a considerable gap in achievement between Black, Hispanic and White students. The 2009-2010 School year data for the state indicates that the pass rate for Math was 74%, 80%, and 90% for Black, Hispanic, and White students respectively, thus revealing a gap of 16% between Black and White Students in math performance. Students with disabilities only had a pass rate of 58%. In addition, Norfolk Public Schools did not meet AYP in Mathematics Performance in three standard NCLB areas: Blacks, Economically Disadvantaged, and Students with Disabilities (Virginia Department of Education Website: Accountability Guide, 2011). Considering that students cannot graduate from high school without passing high-stakes tests in math and that student performance on these tests is critical to the school district in terms of funding, it seems reasonable that research is needed to identify effective strategies to help lower-performing students do better on high-stakes testing.

Existing research suggests that strategies, such as calibration practice, may help these students to improve their metacognition and self-regulation. If they can more effectively monitor their own learning, it may deter them from giving up. By teaching

them self-regulatory behaviors they may improve their academic success. One such indicator of academic success is performance on the SOL exams. Nietfeld, Cao, and Osborne (2006) reported that even modest interventions to improve students' metacognition can improve calibration, performance, and self-efficacy.

### The Relationship between Self-Efficacy and Calibration

Adolescence is the transition from childhood to adulthood, and is generally accepted as occurring between the ages of 11-12 (puberty) and late teens or early twenties (Schunk & Meece, 2006). Adolescence is a particularly formative time, when changes can impact the course of students' lives. Adolescent self-efficacy beliefs are becoming crystallized, which is why early intervention is necessary if self-efficacy beliefs are faulty (Bandura, 2006; Chen & Zimmerman, 2007). Self-efficacy is "an individual's judgments of his or her capabilities to perform given actions" (Schunk, 1991, p. 207). Understanding student self-efficacy beliefs is necessary, especially for math and science educators. If these educators can identify students with poor or faulty self-efficacy beliefs they may help them progress down career paths that they might otherwise overlook, such as engineering and technology.

Self-efficacy is often confused with self-esteem, but self-efficacy is more situation specific than self-esteem. Although a person may have low self-efficacy beliefs in a subject, such as math or science, they may otherwise have high self-esteem (Bong, 2006; Zimmerman & Cleary, 2006). Self-efficacy is important to academic achievement because it is a measure of the student's subjective belief that they can perform a task at a desired level (Bong, 2006). Zimmerman and Cleary suggested that when effects of

cognitive ability are controlled, adolescents' self-efficacy perceptions account for a unique variance in an academic outcome (2006). They refer to a meta-analysis of studies conducted by Multon, Brown, & Lent (1991) that indicated self-efficacy accounted for 14% of the variance in students' academic performance. Pajares (2006) suggests that the variance is even higher at 25%. Students with high self-efficacy are more likely to persist when challenged and effectively use learning strategies (Bandura, 2006; Walker & Greene, 2009; Chen, 2003; Schunk, 1991; Bonk, 2006).

Though there is a relationship between self-efficacy and achievement, it is not perfect. Not all gifted individuals will perform well academically, while some lower-ability students will perform above grade expectations. This may be due to their self-efficacy beliefs. Studies have even shown that students' self-efficacy beliefs may lead to performance above their ability level (Bandura, 1993; as cited in Zimmerman et al., 2006). Students who doubt their efficacy reduce their academic aspirations. In other words, individuals who are more self-efficacious about their skills are more likely to succeed, even when others may have the same skill level (Zimmerman et al., 2006).

It is from this point that it is necessary to proceed with caution, because having high self-efficacy beliefs is not always appropriate. Just having high self-efficacy beliefs is not enough if those self-efficacy beliefs are not grounded in reality. Chen's 2003 research with seventh grade students supported this finding, "...the students relied on the strength and calibration of their pre-performance self-efficacy judgments, rather than on their math performance, to render judgments of effort when solving math problems" (p. 91). If this is true, then to improve a student's performance in a specific subject, it would

be first be necessary to determine their self-efficacy beliefs about their performance in that subject. If their self-efficacy beliefs do not align with their performance, a comparison of their beliefs about their performance with their actual performance over time may be helpful to change persistent but faulty self-efficacy beliefs. This could be accomplished if a self-efficacy measure was taken prior to implementing a calibration strategy over several tests. Students could compare their original self-efficacy beliefs with their calibration accuracy and performance on tests. It is not known if this type of intervention would result in better calibration accuracy, and there is little to no research existing on the relationship between self-efficacy and calibration accuracy. Hacker et al.'s 2008 study involving explanatory styles came close to this, but explanatory styles are different than self-efficacy in that they are more related to what individuals perceive to be the reasons for their performance outcomes rather than their beliefs about what their outcomes will be. A student may believe that she performs well on standardized tests in math, despite evidence to the contrary that she has only barely passed her last two standardized tests. She blames her performance on the tests being tricky, rather than she didn't study for the test enough. This reflects an external explanatory style. An external explanatory style and inappropriately high self-efficacy beliefs may result in the student not preparing adequately for tests.

Prediction or the ability to judge one's own performance is a calibration process closely related to judgments of self-efficacy. Accurate metacognitive monitoring is critical because it provides a basis for the regulation of study (Thiede, Anderson, & Theriault, 2003). Therefore, well-developed metacognition skills such as the ability to

understand cognitive strengths and weaknesses, and self-regulation, are necessary for successful academic functioning (Flavell, 1976). Calibration accuracy is understandably important, but many studies indicate that people are not very accurate when calibrating their performance (Bol & Hacker, 2001; Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010; Kruger & Dunning, 1999). And it is not clear if calibration accuracy is subject to improvement. In Hacker et al.'s 2000 study, high performing college students increased their predictive and postdictive accuracy; however, other studies have failed in attempts to improve calibration accuracy (Bol et al., 2001; Nietfeld, Cao, & Osborne, 2005).

Practice to improve calibration may help students become more accurate at calibrating their own performance. Improved accuracy should ultimately help them learn to understand their own strengths and weaknesses and how to use that knowledge to increase their performance on exams. Learning how to accurately calibrate one's performance is a useful metacognitive and self-regulatory strategy and may especially help lower-achieving students taking high stakes tests. It may prompt them to consider likely outcomes in advance, and therefore take necessary steps to have more control of these outcomes. More productive study and practice would be particularly important in math.

Academic self-regulation requires students to control the factors or conditions affecting their learning by setting goals, determining how to achieve them, and monitoring their progress. Self-regulated learners are usually well calibrated as they are aware of what they do and do not know about a given task. As noted previously, explanatory style may also play a role. Some students view their academic outcomes in

terms of external factors beyond their control, such as luck or teacher bias, and do not learn self-regulatory behaviors. Previous research indicates that relationships exist between calibration and students' explanatory style. Some evidence suggests that overconfidence in predictions may correlate with external attributions, while underconfidence in predictions may correlate with internal attributions (Hacker et al., 2008).

There is a fair bit of research on the relationship between achievement and calibration accuracy. Research indicates that higher performing students tend to be better calibrated than lower performing students (Bol et al., 2001; Bol et al., 2010; Hacker et al., 2000). This is not surprising, considering that higher performing students have likely developed better self-regulation skills where they continually monitor their own performance and adjust their planning and goal setting accordingly. It would next seem logical to expect that higher achievers would have higher self-efficacy beliefs and lower achievers would have lower self-efficacy beliefs, but that is often not the case. Both high and low performers can express overconfidence in their predictions (Bol et al. (2010). This suggests faulty self-efficacy beliefs, but there is little research on the relationship between self-efficacy beliefs and calibration accuracy. Bol and Garner (2011) suggest that low achievers anchor their calibration judgments on optimistic but inaccurate judgments of their ability. More research is needed to determine if a significant relationship exists between these two constructs. Past research has hinted at this relationship (e.g., Hacker et al.'s 2008 study on explanatory style and calibration



accuracy), but the relationship between self-efficacy beliefs and calibration accuracy has not been specifically targeted.

In a study on the self-efficacy beliefs of seventh graders, Chen found that students overestimated their math capabilities, which undermines the predictive power of their self-efficacy judgments (2003). If students do not have accurate self-efficacy awareness of a subject, it seems to follow that their calibration accuracy will be poor. Zimmerman et al. (2006) suggest that introducing adolescents to strategies they control may help them learn to see success and failure in terms of the strategies they use rather than ability, because "...adolescents are often poor at setting goals and anticipating the consequences of various courses of action; as a result, they fail to employ effective task-specific strategies such as preparing for tests" ( p.47).

Calibration strategies may help these students develop academic independence by encouraging them to visualize the results they hope to achieve in advance. It is not enough to just introduce calibration strategies to them because if they do consider the adequacy of their knowledge on upcoming tests, their performance may be skewed when their self-efficacy beliefs result in inaccurate calibration. Therefore it is necessary to have students practice calibration over time to allow them the opportunity to reconcile their results with their predictions.

#### Significance and Purpose of Study

With the exception of the study conducted by Bol et al. in 2010, little research has been done on the effect of calibration on achievement specifically for urban, public middle school students. This is surprising considering that these students are the ones

who may benefit most from strategies that encourage self-reflection and self-regulation, especially in math classes where there are achievement gaps in test results between African American and White subgroups.

High-stakes tests, such as Virginia's Standards of Learning tests and end-of-course exams, are used for accountability purposes that can have harsh consequences for schools, teachers and students. Students in urban environments are more likely to describe high-stakes testing in negative terms more than their peers in suburban schools (Hoffman & Nottis, 2008). One possible reason for this negativity is that younger students may lack the metacognitive skills to accurately calibrate their performance. As a result, they may not prepare for them sufficiently. Students with poor self-regulatory skills must become aware of how their current learning and study strategies influence their learning and outcomes. Researchers have suggested that the lowest performing students lack knowledge both of course content and an awareness of their own knowledge deficits (Dembo et al, 2000; Hacker et al., 2000; Kruger et al., 1999). If students were asked to complete self-efficacy measures, it may help both them and their teachers determine if there were a mismatch between their self-efficacy beliefs and performance. For example, in Bol et al.'s 2010 study, all middle school students were overconfident in their predictions, although lower performing students were more overconfident than higher performing students. Calibration practice may help lower-performing students improve their self-awareness and self-regulatory behaviors by training them to recognize their own knowledge deficits and promote practices related to study and preparation habits.

This research was undertaken primarily to determine if a topical, rather than item level or whole test, calibration strategy would have an impact on calibration accuracy or test performance. Another area of investigation was to determine if self-efficacy beliefs were predictive of calibration accuracy, bias, or performance. There are many studies that use whole test or item level calibration predictions to determine calibration accuracy (Hacker et al., 2000; Hacker et al., 2008). There are also many studies showing a relationship between performance and calibration accuracy (Chen, 2003; Nietfeld et al., 2006). These will be discussed in the next chapter. However, none of these studies used high-stakes math tests as a measure, adding a dimension of motivation for accuracy considering the consequences of failure. Nor were any of these studies implemented in a school largely populated by disadvantaged and minority students, who need to increase their performance on high-stakes math tests or face sanctions. In addition, these students may be most in need of calibration realignment due to faulty self-efficacy beliefs. This study also used a different type of calibration strategy. Students predicted their scores for each sub-topic on a test.

The following research questions were addressed:

1. Does topical calibration practice affect student calibration accuracy or test performance?
2. Do topical calibration practice and achievement level interact to influence prediction accuracy?
3. Does calibration accuracy differ by topic and does accuracy by topic differ by treatment condition?

4. Does self-efficacy of students who practiced by topic predict calibration accuracy and SOL performance?
5. What are the student's beliefs about the efficacy of calibration practice on their accuracy and performance on tests?

#### Overview of Method

A quasi-experimental design was used to address the research questions. One hundred and ten sixth grade students enrolled in regular math in a Norfolk public middle school participated in this study. One sixth grade regular math teacher had students practice a topical calibration strategy for three tests and then the final SOL test. Another sixth grade regular math teacher's students did not practice calibration, but predicted their grades topically on the final SOL exam only. The topical effect on performance was determined by comparing final test scores between students that practiced calibration and those that only predicted scores for the final SOL test. A self-efficacy measure was distributed at the beginning of the semester to students who practiced calibration to determine relationships between self-efficacy, calibration accuracy and performance on the final SOL test.

## CHAPTER II

### LITERATURE REVIEW

#### Overview

This chapter provides a review of the literature relating to calibration. It begins with the types of calibration and the ways calibration is measured. This is followed by a review of literature examining the relationship between calibration and self-efficacy. The relationship between calibration accuracy, achievement, and bias follows. The chapter concludes with a discussion of the effectiveness of calibration strategies aimed at increasing calibration accuracy and performance.

#### Calibration Types and Measures

How we define processes influences the measures we use to assess them and interpret research findings (Schunk, 2008). Therefore, the definitions of calibration and absolute and relative accuracy are necessary. Huff and Nietfeld define calibration as the degree to which one can match their perception of their performance with their actual level of performance (2009). Horgan (1990) simplifies this definition and applies it specifically to students, “the accuracy with which students can predict their own performance” (p.1). Both of these calibration definitions refer to absolute accuracy, as opposed to relative accuracy. According to Hacker and Bol, absolute accuracy is an overall judgment of performance, whereas relative accuracy is the ability to discriminate between items to determine which items are more likely to be correct (2011).

Hacker and Bol assert that both absolute and relative measures of calibration are needed if students are to improve their self-regulation abilities and exert precise control over their learning (2011). In research with 67 undergraduate students who provided absolute metacognitive judgments at the test and item level and relative metacognitive judgments at the item level, the authors found a shared variability between absolute and relative accuracy. They recommended that students make both test-level and item-level judgments of performance to better self-regulate their test preparation.

The decision to use absolute versus relative measures of accuracy is dependent on the goals and context of the study. Generally, if the intent of the research is to compare judgments with actual performance or to determine if accuracy is changed as a result of training or an intervention, absolute accuracy measures are better. If the intent of research is to determine if participants make consistent judgments across items, or can discriminate between items they will do poorly on versus items they will do better on, then a measure of relative accuracy is better (Nietfeld et al., 2006; Schraw, 2009).

An explanation of the differences between the measures is necessary to consider their uses. Absolute accuracy is measured by investigating whether judgments match performance exactly (Maki, Shields, Wheeler, and Zacchilli, 2005). A straightforward method of calculating absolute accuracy is to find the difference between calibrated judgments of correct responses and actual number of correct responses. The closer the number is to zero, the more accurate it is. For example, a student who expects to answer 5 questions out of 10 correctly, and actually does answer 5 questions correctly has an accuracy score of 0, calculated as their predicted score (5) minus their actual score (5).

An accuracy score of 0 means the student is perfectly calibrated. A student who expects to answer 5 problems correctly out of 10 and actually answers 3 problems correctly has an accuracy of 5 minus 3, which is 2.

Whereas absolute accuracy is the difference between a calibration judgment and the actual score, a bias calculation is used to determine the direction of the miscalibration. A positive number reflects overconfidence, while a negative number reflects underconfidence. For example, if a student predicted they would answer 8 out of 10 problems correctly, and their actual grade was a 6 out of 10, then their bias score is their predicted score (8) minus their actual score (6), which is 2, reflecting overconfidence. In contrast, a student who expects to answer 6 questions correctly out of 10 and actually answers 8 correctly has a bias score of 6 minus 8, which is -2, reflecting underconfidence.

Bol, Riggs, Hacker, Dickerson, and Nunnery's 2010 study with 77 middle school math students calculated absolute accuracy and bias scores in this manner. Students were asked to predict how many problems out of 50 they would solve correctly on the state's standardized, end-of-year test. Participants came from 2 regular classes and 2 honors classes. Students in the regular classes had mean absolute prediction accuracy of 9.5, whereas students in the honors classes had a mean absolute prediction accuracy of 5.6. Students in the regular classes had a mean prediction bias score 7.3, and the honors' students had a mean bias prediction score of 1.3. Both groups of students were overconfident, as indicated by the positive bias numbers, but students in the honors group were only slightly overconfident when compared to students in the regular classes.

When making item level judgments, students are often asked to express their confidence, either before attempting to solve the problem, a predictive judgment, or after attempting to solve the problem, a retrospective judgment. In 2003, Chen calculated absolute accuracy bias at the item level a bit differently than explained above. She asked 107 middle school students to rate their confidence in individual math problems before they attempted to solve each problem. A performance rating was assigned based on whether the answer was correct or incorrect. Correct answers received a score of 8 and incorrect answers received a score of 1. A confidence rating was assigned using a scale from 1 to 8. Students were asked to answer the question, "How confident are you about solving this math problem correctly?" A rating of 1 represented 'not at all confident', and 8 represented 'completely confident'. Students could pick any number between 1 and 8 to express their confidence. The difference between the performance score (1 or 8), and the confidence score (ranging from 1 to 8) was the bias. For example, a student who reported complete confidence in solving the problem (8), but got it wrong (1) had a bias score of +7 (the difference of  $8-1$ ). The positive number reflects overconfidence. A student who was only mildly confident they would solve the problem might have assigned a confidence rating of 6. If they solved it correctly they would get a performance score of 8, so their bias score would be -2 (the difference of  $6-8$ ). The negative number reflects underconfidence. Absolute accuracy was then determined by subtracting the absolute (unsigned) value of each bias score from 7. Using the last example, the bias score of -2 would change to the unsigned value of 2, and be subtracted from 7, resulting in 5 ( $7-2$ ). In this method of calculation, a 0 represents complete inaccuracy, whereas a 7 represents



complete accuracy. A 5 would therefore suggest reasonable accuracy. Chen's performance measures had three levels of difficulty (easy, moderate, and difficult), and gender was employed an independent variable. The combined average accuracy was 3.78 for girls and 3.74 for boys. This suggests that boys and girls were inaccurate to the same extent, as their accuracy was at the midpoint between completely inaccurate and completely accurate. The combined mean bias scores for girls were 2.46, while the combined mean bias scores for boys was 2.39. Positive numbers for bias scores reflect overconfidence, so both boys and girls were slightly overconfident in their accuracy judgments.

Relative accuracy is calculated by correlating confidence judgments for individual items. Maki et al. (2005) and Hacker et al. (2011) suggest the non-parametric gamma to measure relative accuracy, although they both acknowledged that there are questions about the reliability of gamma. Gamma shows whether items that receive higher confidence judgments (calibration) scores result in higher performance, and whether items that receive lower confidence judgments result in lower performance (Maki et al., 2005).

Maki et al. collected relative accuracy data from 159 general psychology class college students. They used the gamma correlation to measure whether high judgments of confidence produced high percentages of correct answers. The gamma ranges from a -1.0 showing a perfect negative relationship, to +1.0, showing a perfect positive relationship. Gammas were calculated for each student using posttest confidence judgments of percent correct and their actual percent correct for 3 conditions related to reading texts. One

condition had hard texts, a second had featured the same hard texts that were revised to be slightly easier, and the last condition had a mix of both kinds of texts. The mean gamma correlation for the hard text condition was .57; it was .36 for the revised text, and .42 for the mixed texts. In other words, all of the confidence judgment gammas were significantly greater than zero. The gamma was higher for harder texts, suggesting greater accuracy with increasing difficulty.

Hacker and Bol (2011) also used a discrimination index in their study to assess the degree to which calibration judgments for correct answers are distinguished from confidence judgments for incorrect items. Schraw explains that the discrimination index adds another dimension to metacognitive monitoring that bias measures miss because it monitors confidence across items rather than performance on a specific item (2009). To obtain the discrimination index number, the mean confidence for incorrect answers is subtracted from the mean confidence for correct answers. If the resulting value is positive, confidence is higher for correct items. A negative value means that confidence is lower on correct items.

The authors (Hacker & Bol) reported that the discrimination index showed moderate to strong correlations with gamma, and concluded that these two measures tap into similar metacognitive monitoring processes. Participants were 57 college students enrolled in an educational psychology class. The students were asked to postdict how well they did on each problem immediately after solving it by answering the question, "How confident are you that your answer is correct?" They were provided with a line graph, with 0% on one end, 100% on the other, and increments of 20, 40, 60, and 80

labeled on the scale. , Students were instructed to select a number on the graph between 1 and 100 that indicated their confidence level, with 0 being not at all confident and 100 being completely confident. Analysis revealed that students were correctly adjusting their confidence judgments for correct and incorrect answers; mean confidence levels for correct answers were 82.25 for true/false questions, 83.69 for multiple choice questions requiring recall of information (MC-Lo), and 83.07 for multiple choice questions requiring analysis (MC-Hi). The confidence judgments for incorrect answers were 68.60, 64.87, and 70.15 for true/false, low difficulty multiple choice, and high difficulty multiple choice items, respectively.

When students predict or postdict their overall scores on a task, it is generally referred to as a global judgment. Predictions or postdictions at the item level are called local judgments. The literature is mixed as to whether global or local judgments are more accurate, and if either is better than the other at improving students' metacognition. Intuitively, it would seem likely that local judgments of comparison would be more accurate as the individual only has to consider the performance on a single item rather than many.

Liberman's 2004 research refutes this idea. Her research with 134 Israeli business students investigated the accuracy bias differences between global and local postdictions. She conducted three studies; participants were asked to rate their confidence after answering each question. The questions were similar to true/false questions; participants were given the names of two cities in each question and asked to identify which city had the larger population. This meant that students had at least a 50% probability of

answering the question correctly even if they guessed. In the first study, half of the students were placed in the restricted condition and instructed before beginning that their global estimates should be 50% or higher, because even if they were to guess, 50% of their answers would be correct. They were told that the lower boundary of their postdictions should therefore be 50%. In the unrestricted condition, the other half of the students were not given these instructions. Students in both conditions were more accurate at the global level than at the local level. Students in the unrestricted condition made global postdictions that were within 5.5 mean points of their actual scores, while their local postdictions were much farther apart at 14.2 mean points of their actual scores. The same was true to a lesser extent for the restricted condition; students made global postdictions that were 12.5 mean points from their actual scores and local postdictions that were 18.5 mean points from their actual scores. Students only expressed underconfidence in the unrestricted global condition; they were 5.5 points under their actual scores. The remaining scores all reflected overconfidence, most notably the local postdictions which were 14.2 and 18.5 mean points over the actual scores in the unrestricted and restricted conditions respectively.

The author did two more studies that included restricted and unrestricted conditions. In the second study, participants in the unrestricted condition were within 4.1 points of their actual scores in their mean local postdictions, while they were within 8.1 points of their actual scores in their mean global postdictions. In other words, participants in the unrestricted condition were more accurate with their local postdictions than their global postdictions. In the restricted condition, scores matched the first study more

closely. Students were again more accurate with their global postdictions, which were within 4.4 points of their actual mean scores than their local postdictions, at 9.3 points from their mean scores.

The third study resulted in global postdiction scores that were more accurate than local postdiction scores in both conditions. The unrestricted group had mean global postdiction scores that were within 6.6 points of their mean actual scores, while the local postdictions were 7.9 mean points over their actual score. The restricted group had global postdiction scores that were 3.4 mean points over their actual mean scores, while their local scores were 11.9 mean points over their actual mean scores.

Students were overconfident in their local postdictions in all three studies, whereas they were underconfident in the global unrestricted conditions and overconfident in the global restricted conditions. The author interpreted these results to suggest that participants would not consider the likelihood of getting at least 50% of their answers correct from random guessing without being told to do so, unless they were in the restricted conditions. Students who were told that their lower boundary should be 50% ended up with postdiction scores that expressed overconfidence in all three conditions. Students who did not receive that instruction were underconfident in their postdictions in all three conditions.

Whether or not local strategies are better at improving calibration accuracy, an important consideration is that local calibration may be problematic if used during test-taking. Grabe and Flannery (2010) found that poorly performing students are sometimes motivated more by completing a task quickly rather than gaining an understanding or

learning. Their study evaluated the use of online study questions that students could answer prior to taking an examination to help them practice for the exam. Answering the online practice questions was voluntary, although students could earn up to 1% of their course grade by accumulating up to 3 points towards each examination. Students accumulated these points by answering the online practice questions for each chapter, and each examination covered three chapters. This resulted in the 3 maximum points towards each examination, for a total of 9 points possible across three examinations. In addition, students had to assign a local confidence rating after answering each question that affected their point total. Prediction accuracy was rewarded with higher possible point values, whether or not the accuracy was in answering a question correctly or incorrectly. However, after the first test, the researchers found that they were unable to collect meaningful data because the lower-performing students had a much lower participation rate for the pre-test study questions for the second and third tests, resulting in skewed data. The researchers theorized that students “who struggled with the system and who may have found prediction difficult simply stopped trying, making it less likely that the hypothesized relationship between prediction accuracy and performance across all participants would be demonstrated” (p. 469). This has negative implications for attempting to use local calibration practices during testing, as the goal is to help students consider their performance better. The results of this study seem to suggest that rather than increasing their testing awareness, local calibration may lead to increased frustration, especially for lower-performing students. The researchers still support the use

of local calibration practices with modifications in the future, as the results are finer-grained than global pre- and postdictions (Grabe & Flannery, 2010).

Although not relating specifically to the relationship between global vs. local calibration and performance, a study conducted by Grimes (2002) added an interesting twist to the global vs. local calibration question. Grimes added a pretest instrument that included a list of 40 economic concepts for 253 macroeconomics students, asking them to indicate with a yes or no answer which concepts they expected to be tested on in the midterm exam. Grimes did this to find out what material the students expected to be on the test, because knowledge and awareness of test subject matter seem necessary for students to determine their study strategies and behavior (2002). As Horgan indicated, students who are underconfident may study much more than necessary, while students who are overconfident and think they know more than they do will not study enough (1990). In addition to this pretest survey of concepts, Grimes asked students to answer two questions, 'What score (between 0 and 100) do you expect to receive on this examination?' and 'Compared to the first examination, do you think you will do better or worse on this examination?' Students were asked both of these questions forty-eight hours prior to the exam, and again just before the exam, when they answered the survey of concepts, and after the test, when they completed the survey of concepts again.

Grimes found a statistically significant inverse correlation between expectation of concepts and predictive calibration. The more incorrectly expected concepts, whether the concepts were ones the students expected to be on the test and were not (False Positives), or the ones they didn't expect to be on the test and were (False Negatives), were named

by the student, the less accurate the student's predictive calibration. This makes sense, because students need to understand what they will be tested on in order understand what they need to do for preparation, for example spend more or less time studying.

Grimes' study was particularly useful in that it emphasizes the need for strategies that go beyond global and local predictions. Adding the concept survey to the postdiction helped explain to an extent why students are miscalibrated, and in this case overconfident; they were incorrect in their assumptions about the material to be tested. Perhaps that is why the students did not perform as well as they may have expected or wanted. If this is true, than it may benefit students to ask them to identify how well they expect to perform on each topic of the test by providing them with a detailed list of concepts prior to the exam. Ideally, providing students with a list of concepts they would be tested on would allow them time to consider the topics on the exam and plan their study accordingly. If a person can accurately discriminate between better learned and less learned material, it may encourage them to regulate their study time more effectively (Thiede et al., 2003). However, being able to discriminate between better and less known material is a metacognitive ability that some students lack due to faulty self-efficacy beliefs.

#### The Relationship between Calibration and Self-Efficacy

Kruger and Dunning suggest that, students may be unaware they are unskilled in a subject, and thus be surprised when they perform poorly (1999). These feelings of competence or incompetence in a domain are known as self-efficacy beliefs. Self-efficacy beliefs are domain specific, such as a person's belief in their ability to solve algebra



problems. In contrast, self-concept is more of a global and general self-perception. (Schunk et al, 2006).

Nietfeld et al. argue that if self-monitoring is a part of other regulatory processes, such as planning, it must also be connected with self-efficacy (2006). The relationship between the ability to accurately calibrate one's judgment and self-efficacy is not clear. One would expect a positive relationship. Students with high self-efficacy would be better at calibrating their performance accurately than students with low self-efficacy. Furthermore, research has not investigated if self-efficacy has an effect on the stability of calibration accuracy over time. For example, are students with higher self-efficacy more likely to improve their calibration accuracy with practice, as opposed to students with lower self-efficacy, who may not even try? And does self-efficacy correlate positively with performance; will students with high self-efficacy also perform better than students with low self-efficacy? Research does not consistently support positive relationship among self-efficacy, calibration accuracy, and performance. Over confidence seems to be a common theme in student predictions of their test scores regardless of subsequent performance (Bol et al. 2010; Chen, 2003; Klassen, 2006). Schunk and Meece argue that strategies to improve self-efficacy should involve ways to keep students informed of their progress in learning (2006). A calibration strategy that requires students to repeatedly calibrate their performances, such as predicting their test scores over time, meets this condition because students are able to see their progress over the semester as they track their calibration accuracy and test scores. However, Nietfeld et al. (2006) noted that there

is not much research that has examined the relationship between self-efficacy monitoring accuracy outcomes.

It seems logical to assume that if students predict their performance across several tests, it keeps them informed of their progress by providing them with a method of self-monitoring. If they are able to improve their calibration accuracy, it may lead to more accurate perceptions of self-efficacy. For example, a student who is consistently overconfident in their predictions of performance on a test may realize, after inaccurately calibrating their performance and failing the task, that their judgments are faulty. As the student realizes that they are not doing as well as they expected, they may adjust their self-efficacy beliefs and behavior accordingly, as they come to understand the need for more preparation and study.

Research has consistently shown that many students, particularly lower achieving students, overestimate their performance (Chen, 2003; Bol et al., 2010, Kruger et al., 1999), indicating faulty calibration processes. Chen (2003) asserts that students' ability to accurately calibrate is pedagogically important, because poor calibration may negatively influence their self-efficacy, "Many at-risk students have reported unrealistically high self-efficacy and an unwillingness to change their study methods because of this overconfidence (p. 81)." Ramdass suggests that self-efficacy is related to calibration because a person's judgment of performing a particular task has to be accurate. He suggests that misjudgments of personal efficacy towards over or underconfidence in a task can have detrimental effects on the outcome of that task (2009).

Klassen's 2006 study with learning disabled (LD) and not learning disabled (NLD) students supported Chen's idea. The study revealed that prediction and self-efficacy scores were highly correlated. He found that significant mismatches between self-efficacy beliefs and performance were linked with poor performance. Students with LDs (and the lowest-achieving students across disability categories) "grossly miscalculated" their ability to compete literacy tasks—their optimistic beliefs did not lead to superior performance" (p.195). While Bandura (1986) postulated that the most functional efficacy judgments are those that slightly exceed one's actual capability, most studies, like Klassen's, indicate that lower achieving students are too overconfident (Bol et al., 2001; Bol et al., 2005). Although overconfidence may sometimes be helpful, it may also be harmful for lower-achieving students, who may rely on overly positive self-efficacy beliefs and not prepare appropriately as a result.

The problem of overconfidence resulting from erroneous self-efficacy beliefs was demonstrated in Bol et al.'s research with seventy-seven sixth graders (2010). Themes emerged from the qualitative data supporting the mismatch between beliefs and performance. Even though students were aware of upcoming SOL math tests, it was common for many of them to discount the need to study because they believed they were either good at that subject or not. Similarly, some thought that they would achieve a certain score because they felt good or bad about the test, or believed that they had studied enough to pass the test. In other words, many students calibrated their performance outcome expectations on faulty self-efficacy beliefs about math. This was

evidently a problem because only one percent of participants expected they would fail the exam, but forty percent actually did fail (Bol et al., 2010).

One of the goals of this study is to provide a calibration strategy that offers students practice predicting their test scores over time, so that they learn to monitor their own self-assessment skills. This should help them align their self-efficacy beliefs with their performance and hopefully, use this knowledge to guide them towards taking the necessary steps to improve their performance. Kruger and Dunning's research supports this prediction, even as they point out the logical paradox. Once students gain the metacognitive skills to recognize their own incompetence, they are often no longer incompetent (1999).

#### The Relationship between Calibration Accuracy, Achievement Level, and Bias

There is a fair bit of research supporting the relationship between calibration accuracy and achievement. As discussed in the previous section, Kruger and Dunning (1999) assert that "...incompetent individuals have more difficulty recognizing their true level of ability than do more competent individuals and that a lack of metacognitive skills may underlie this deficiency" (p.1122). The authors suggest that people who actually have knowledge in a specific domain are more likely to recognize what they do not know about that subject, and therefore are better able to calibrate their judgments than those lacking knowledge in that domain. The literature on calibration supports this idea, repeatedly. As Schraw acknowledged in 1994, successful learners already use different cognitive and metacognitive skills to improve their learning. Maki et al. (2005) point out that in general, people who have less knowledge in a specific domain tend to largely

overestimate their performance, while people who have a greater knowledge slightly underestimate their performance.

Koku and Qureshi's 2004 research supports these findings; they assert that high performing students are more likely to recognize the extent and limitation of their knowledge. Low performing students have limited insight into their performance. Ninety-one students majoring in business were assigned to one of three treatment groups. All students took the same multiple choice final examinations. All students were also asked to provide a postdictive confidence judgment at the item level about how certain they were that they answered correctly. The authors divided the students into high, average, and low groups based on their exam performance. High performing students' mean probability judgment was 83% while their actual mean score was 80%, indicating slight overconfidence of 3%. Average students' mean probability judgment was 78%, while their actual mean score was 68%, resulting in a mean overconfidence of 10%. Low-performing students mean probability judgment was 73% while their actual mean score was 51%, representing an overconfidence of 22%. This illustrates clearly that student overconfidence is more pronounced with lower performance.

Research consistently reinforces the theory that more successful students seem to underestimate their knowledge and performance on tests, or marginally overestimate; however, less successful students overestimate their knowledge and performance on tests, sometimes grossly so (Bol et al. 2005; Hacker et al., 2000; Koku et al., 2004; Kruger et al., 1999). This seems to be a contradiction as one would expect lower achieving students to express underconfidence rather than overconfidence. There are several

theories as to why this happens. Much of the literature about explanatory style suggests that this contradiction may occur as a result of lower-achieving students' protection of self-worth and image or their desire to appear as good students (Butler & Winne., 1995; Bol, et al. 2001; Bol et al., 2005; Dembo et al., 2000). Koku et al. contend that lower achieving students deliberate less among multiple choice answers if they believe they already know the correct answer. This lack of deliberation results in inadequate consideration of all of the choices (2004). These students may feel that they are more familiar with the content than they are, and their bias leads them to their first choice of answer without careful consideration. Hacker et al. suggest that lower-performing students base their poor performance on tests to external factors, such as the teacher's poor teaching or the test itself being tricky or too hard, rather than on past exam performance (2000). Bol et al.'s 2001 research mirrored the 2000 study, but the authors suggested that low achieving students based their predictions on global self-concepts of academic ability rather than past exam performance.

The various factors that students represent as causes for why they do or do not do well on tests are known as attributions. Zimmerman and Cleary (2006) among others have found a relationship between students' causal attributions and their self-efficacy beliefs (Bol et al. 2005; Hacker et al., 2008). For example, students may have a mastery orientation and attribute successes and failures to internally controllable factors, such as effort. Other students may attribute their successes and failure to external factors, such as luck. Students who exhibit higher self-efficacy tend to attribute their performance outcomes to internal controllable factors, while less efficacious students tend to attribute

performance to uncontrollable external factors (Zimmerman, et al., 2006). Hacker et al.'s (2008) research linking trends in their findings suggested that calibration and students' attributions supported these relationships. Overconfidence in predictions may correlate with external attributions, yet underconfidence in predictions may correlate with internal attributions. Bol et al.'s 2005 study with undergraduate college students supported the theory that explanatory style is linked with calibration. Furthermore, higher achieving students were more accurate and underconfident in predictions, while lower achieving students were overconfident and inaccurate.

In a study conducted with school-age children, the results were similar to those of the college students. Bol et al.'s 2010 study was with 77 urban, middle school students who were asked to globally predict and postdict their test scores on SOL tests for 6th grade mathematics. Two regular and two honor's math classes participated, and both prediction and postdiction values show that higher achieving students were more accurate. Higher achievers were generally within 4 points of their actual score, while the lower achievers were within 10 points. While this evidence clearly suggests that a relationship exists between performance and calibration accuracy, it is interesting to note that age appears to be a factor in the direction of calibration bias. Schraw (1994) suggested "...older learners are accurate judges of their monitoring ability even though they lack the on-line regulatory skills necessary to monitor effectively" (p. 153). College students' bias appears to be more related to their performance. As noted, college students who are high-performers tend towards underestimation, yet lower performers tend towards overestimation (Bol et al., 2001; Bol et al., 2005; Hacker et al., 2008). This is

does not always seem true with the younger students who tend towards overconfidence regardless of achievement level (Boekaerts & Rozendaal, 2010; Bol et al., 2010; Chen, 2003).

Although research with younger students suggests overconfidence bias for both high and low performers, Boekaerts and Rozendaal's (2010) study showed that overconfidence was more pronounced with difficult problems. The authors presented the fifth graders with math application problems (word problems) and computation problems and asked them to measure their local level of confidence both before and after solving individual problems. Student bias was analyzed by problem type (application-harder or computation-easier). Students tended to overestimate their performance on the more difficult application problems than on the computation problems. Flannelly's (2001) study provides further support. In that study, nursing students who performed poorly on tests were more overconfident about their answers to hard test questions than students who performed better. This is so widespread a phenomenon that it has been named the hard-easy effect (Hacker et al., 2009). This implies that a relationship exists between the difficulty of the material and calibration accuracy. This has been documented repeatedly (Flannelly, 2001; Nietfeld et al., 2005; Winne & Jamieson-Noel, 2002). The phenomenon makes sense considering that higher performing students tend to be better calibrated than students who are lower performers. Higher performing students may be more aware of the difficulty of the material to be solved and their own ability to correctly solve it, thereby making them better calibrated and less overconfident than lower performing students. The lower performers may find the material more difficult, causing



them to calibrate less accurately and more overconfidently since they are unaware of the parameters of their own knowledge deficits (Kruger et al., 1999).

The effects of gender on bias have also been investigated, with somewhat mixed results. Boekaerts and Rozendaal (2010) found that boys overestimated their performance more than girls, although both boys and girls expressed overconfidence. These results support Chen's 2003 study with seventh grade math students. Chen found that boys tended to evaluate their math performance more favorably than girls, and that all of the students overestimated their math capabilities. However, the calibration accuracy and bias of both genders was comparable, so gender was excluded from Chen's final path analysis model (p.89).

These researchers (Chen, 2003; Boekaerts & Rozendaal, 2010) and others call for research to identify how teachers can help their students become better self-regulated learners and more accurate calibrators. Further investigation is needed, for both high and low achieving students, on ways to improve calibration accuracy. It seems logical though to be more concerned with the lowest achieving students. They are often inaccurate and overconfident when predicting their performance and most in need of effective interventions (Bol, et al., 2005; Hacker & Bol, 2004; Pajares & Miller, 1997). They are also the students most in danger of failing high stakes tests

#### Calibration Strategies

It is still unclear whether calibration accuracy can be improved, despite numerous studies attempting to do so (e.g., Bol & Hacker, 2001; Bol et al., 2005; Hacker et al., 2000; Neitfeld et al., 2006). Lin and Zabrocky (1998) extensively reviewed the literature

on the calibration of comprehension. They developed a table summarizing 34 studies on how subject, task, and text variables influenced calibration. Research into calibration has continued to evolve and expand since then, encompassing many variables related to calibration, most notably the ability to improve calibration accuracy, and its relationship to self-efficacy and performance. Hacker et al. (2009) summarized the characteristics and major findings of 12 calibration studies in classroom contexts, excluding studies done in laboratory settings. The remaining portion of this chapter will review calibration intervention studies that were focused on improving calibration accuracy, decreasing calibration bias, and ultimately improving test performance.

Nietfeld, et al.'s 2006 research focused on an intervention consisting of monitoring exercises and feedback to see if calibration accuracy and student performance outcomes could be improved, and if a change in calibration could account for changes in self-efficacy over a semester. Eighty-four college students in two different classes took an educational psychology self-efficacy inventory during the first and last class of the semester. This was used to measure changes in self-efficacy over the semester. Students in one class were in the treatment condition, while students in the second class were in the comparison condition. Students in both conditions recorded confidence judgments at the local level for each of four tests throughout the semester. Students in the treatment condition also completed monitoring worksheets at the end of each non-test day class, after the introductory class. The worksheets asked students to do four things: 1) rate their understanding of the day's content, 2) explain which concepts they found difficult to understand, 3) explain what they could do to improve their understanding of difficult

concepts, and 4) answer three multiple choice review questions from the day, followed by a confidence judgment about their accuracy on a scale from 0% Accurate to 100% Accurate for each review question. The review questions were answered before class ended so students could compare their judgments with their actual performance. Students kept their monitoring worksheets and were encouraged to use them to guide their study and review process throughout the semester. They were also provided with feedback on their overall calibration and bias scores from each test, with instructions on how to interpret their scores. Students in the comparison condition were given the opportunity to self-generate feedback between their confidence judgments and performance after each test, but did not complete monitoring worksheets or receive calibration and bias scores.

Calibration accuracy improved for students in the treatment condition. Initial calibration accuracy ratings were similar for both conditions, but for each of the following three tests, students in the treatment conditions were significantly better calibrated than students in the comparison condition.

Students in the treatment condition also performed better on three of the four tests across the semester. The first test only showed a mean difference between the two groups of 3 points, but the treatment condition students outperformed the comparison group students by 9 points on the second test, 15 points on the third test, and 4 points on the final test. Follow-up analyses indicated that their performance on tests 2, 3, and 4 were significantly different. The scores for the last three tests showed that students in the treatment condition were consistently one standard deviation better than those in the comparison condition.

The authors suggest future research in the fields of mathematics or chemistry, where knowledge builds on previous knowledge, to test whether intervention effects continue over time. They also suggest further research in developing the metacognitive skills of students in K-12 settings to see if younger students may be helped to develop metacognitive abilities enhance performance (Nietfeld et al, 2006).

A study with younger students was conducted by Brookhart, Andolina, Zuza, and Furman in 2006. The results indicated that it is possible to increase calibration accuracy in younger students. Their study was in a suburban setting with students who were learning multiplication tables. They were asked to predict how they would do on weekly tests, and graph their predictions. They were also required to graph their actual results on the same chart. At the end of 10 weeks, the researchers found that student predictions became more accurate with time, although student prediction accuracy varied widely. The authors concluded that self-assessment strategies were beneficial to these students, although they also found it was context dependent.

A study with fifth-grade students conducted by Huff and Nietfeld (2009) also indicated that calibration accuracy improvement is possible with this population. The researchers compared process-oriented monitoring strategies with response-oriented activities, and found an increase in calibration accuracy but not in performance. They describe process-oriented approaches as strategies that teach students to monitor their comprehension during reading, while response-oriented procedures encourage students to consistently calibrate their retrospective confidence judgments with performance.

The researchers used four groups. The first group used the process-oriented approach. This involved teaching students to use comprehension 'fix-up' strategies while they were reading and was identified as the comprehension monitoring training condition. Students in this group received reading passages, with places marked for them to consider their comprehension of each passage, but without instruction on how to reflect. The next group combined the process and response approaches and was called comprehension monitoring and monitoring accuracy training. Students were provided with instruction on how to monitor their calibration in addition to the comprehension monitoring training described above. Students were additionally asked to reflect on 3 questions about their confidence ratings. Class discussion followed, and students were encouraged to try to improve their confidence judgment accuracy. The third condition was the control condition. Students participated in reading the passages and were provided with the answers at the end of the session. They did not receive any instruction, nor did they provide confidence judgments. The fourth condition was the no intervention condition. Students read the passages, and provided confidence judgments for each item without any instruction.

Huff and Nietfeld (2009) found significant results relating to calibration accuracy and confidence judgments, but not on performance. Students in both the comprehension monitoring training and comprehension monitoring and monitoring accuracy training conditions showed significant increases in calibration accuracy and confidence; whereas, the comparison condition did not change from the beginning in accuracy or confidence.

While student performance improved from the beginning in all four groups, the interaction between the treatment groups and performance was not significant.

Studies with older students, such as Hacker et al.'s 2000 study with 99 introductory educational psychology students showed that the highest achieving college students' calibration accuracy improved with practice. Students calibrated their answers globally on three multiple-choice tests, and were also frequently encouraged to reflect on how self-assessment is related to performance. The authors found that low-performing students showed poor prediction accuracy on the first exam and did not improve by the end of the course, whereas, high-performing students showed poor prediction accuracy on the first test but showed increases in predictive accuracy by the end of the course.

In contrast, Hacker et al.'s 2008 study partially supported the idea that improvements in calibration accuracy can occur, but only in one condition. One hundred and thirty-seven students enrolled in an introductory educational psychology course participated by making global predictions for three exams throughout the semester. Students were separated into four conditions; 1) those that received extrinsic incentives, 2) those that used reflection strategies, 3) those that received extrinsic incentives plus used reflection strategies, and 4) those that did not receive incentives or use reflection strategies. Students in the two groups that received extrinsic incentives were told they would receive from one to four additional points on the tests, depending on the accuracy of their predictions. Students in the two groups asked to do self-reflection answered questions such as "How would you explain any discrepancy between how well you

thought you would do on the first exam and how well you actually did?" Students in the comparison group did not receive extrinsic incentives and were not asked to self-reflect.

Results showed no significant group differences for calibration accuracy across the tests except for the extrinsic rewards group. The authors asserted that prediction accuracy was 'remarkably stable' (p. 26). However, a relationship between calibration and performance was again indicated; higher-performing students were 94% accurate in their predictions and postdictions, compared to lower-achieving students, who were 86-88% accurate in predictions and postdictions. In addition, an analysis of interactions amongst the lower achieving students did indicate that those in the extrinsic incentives condition had significantly higher accuracy scores than those students in the other groups.

Bol et al.'s 2005 study also indicated that postdiction may improve with practice, but not prediction. College students in an introductory education class separated students into overt and covert practice calibration groups. Students in the overt groups were asked to enter their prediction and postdiction estimates online for five quizzes leading up to a final exam. Students in the covert condition did not practice calibration overtly by entering their estimate information, but the researchers argued that they could have been making covert predictions. Students in both conditions were asked to make global predictions and postdictions of their performance on the final exam.

Results indicated that overt practice manipulation did not have a significant impact on students' prediction or postdiction accuracy. However, the authors did find a linear trend in the data for postdiction accuracy showing increasing accuracy with repeated practice. They posited that as postdictions are made after the test; some of the

guesswork about test specifics is resolved, resulting in greater postdiction compared to prediction accuracy.

Flannelly (2001) found that a strategy that provides even indirect feedback to students on the accuracy of their confidence and performance slightly decreases overconfidence bias. Nursing students were assigned to an experimental condition and a control condition. Both groups of students participated in a 2- hour pre-test review session one week prior to the post course test. Students in the experimental condition were given a practice test, as well as the answer key as feedback on their ability to pass the test during the review session. However, these students were not provided with direct feedback from the teacher. Students in the control group did not receive the practice test. Regardless of experimental condition, students who performed lower on the test were overconfident that their answers to hard questions were correct. In contrast, students who performed better were underconfident about their answers on hard questions. However, students who received the indirect feedback from the practice test exhibited less overconfidence on hard questions. The author theorized that students who received feedback decreased their confidence that they were right on hard questions. The author suggested future research using more direct forms of feedback prior to test taking.

Although the feedback strategy just discussed was not as effective as may have been hoped, it is a somewhat promising strategy that should be investigated further. The intent of the current study was also to provide feedback prior to test taking. Students were introduced to a topical calibration strategy to help prepare them for the high stakes tests. Instead of making predictions for each item or for the entire test, students practiced



calibration for topics. It was hoped that a significant increase in calibration accuracy and decrease in bias would be obtained.

### Rationale for Study

Research is mixed as to whether calibration accuracy can be improved, and many authors have called for further research into calibration (Bol et al., 2005; Cleary, 2009; Hacker et al., 2008; Nietfeld et al., 2006; Pajares et al., 1997). Recall that according to Pajares (2006), self-regulatory habits are developed early, however, much of this calibration research was primarily with college age students (Bol et al., 2001; Bol, et al., 2005; Hacker et al. 2008; Nietfeld et al., 2006). Schraw suggests that metacognitive knowledge appears to be an important constraint on adult cognition, but it is still uncertain if the same is true of younger populations (1994). Roebbers, Schmid, and Roderer (2009) found that although monitoring and control processes in test-taking are developed in younger test-takers by the age of 9, these skills are still open to improvement, especially in their monitoring of their incorrect answers. Older students have had years of schooling to solidify their beliefs about their own abilities This may affect their ability to change their perceptions about their own self-efficacy and resulting calibration judgments. There are a limited number of studies that have been conducted with adolescents to see if their calibration judgments could be improved with calibration practice. The present study will help address this gap.

Research is needed to further examine the link between calibration and performance, as well as what calibration strategies best promote achievement. While other studies have had students predict performance by individual item (Ramdass, 2009;

Nietfeld et al., 2006; Huff et al., 2009; Chen, 2003) or by global test performance (Bol et al., 2001; Hacker et al., 2008), this study included a strategy that required students to predict how well they did test topic. This is less time-consuming than item-by-item calibration, and potentially more effective than global test calibration. In addition, SOL tests provide course blueprints indicating specific topics that students were tested on, allowing for a topical calibration condition.

In their calibration research, (2012), Bol et al. emphasize the need for effective ways to increase student achievement because of the emphasis placed on testing scores. The current study focused on test calibration practice for the high stakes test. The high stakes and task authenticity should have motivated students to seriously consider the strategies. As Pajares and Graham (1999) concluded, the predictive power of efficacy assessments is maximized when actual high-stakes tasks are presented rather than simulated experiences. Their study did use end-of-unit exams that were factored in for grade promotion; however, there were very few studies, relating to calibration and high-stakes testing like the SOLs. Bol et al.'s 2010 study is an exception, but more research is warranted.

Research also shows that students who are able to accurately calibrate their performance tend to be higher achievers. Underachieving students are often not aware of their metacognitive deficiencies. Klassen (2006) suggests that approaches are needed for Learning Disabled (LD) students that foster their self-awareness and self-regulation as a way to improve their calibration accuracy and subsequent performance. More calibration research is needed, especially for under-performing students facing high-stakes tests, who

are often miscalibrated and overconfident. Math is especially important both for our students' futures and to keep the United States globally competitive. The purpose of the current study was to introduce a calibration strategy to math students preparing for a high stakes test to determine if it had a positive effect on their calibration accuracy and test performance, and to see if their bias (overconfidence or underconfidence) may be diminished.

### Research Questions and Hypotheses

This research investigated whether a topical calibration strategy would have an effect on calibration accuracy and bias, and test performance. Another important question was whether accuracy varies by topic, and if so, if there was there an interaction between treatment group and topical accuracy. Another question was whether there was a predictive relationship between self-efficacy and achievement, and self-efficacy and calibration accuracy. Whether or not students find calibration strategies helpful will also be investigated. More specifically, the following research questions were addressed:

1. Does topical calibration practice affect student calibration accuracy or test performance?
2. Do topical calibration practice and achievement level interact to influence prediction accuracy?
3. Does calibration accuracy differ by topic and does accuracy by topic differ by treatment condition?

4. Does self-efficacy of students who practiced by topic predict calibration accuracy and SOL performance?
5. What are the student's beliefs about the efficacy of calibration practice on their accuracy and performance on tests?

It was hypothesized that students who practice calibrating their scores prior to the SOL would have better calibration accuracy than students who only predicted their scores for the final SOL exam. In addition, it was hypothesized that students who practiced calibration would do better on their tests than students who only made predictions on the final SOL test, as they had time to compare their judgments with their outcomes over several tests. It was additionally expected that students who performed better on the SOL test would be more accurate in their predictions than lower performing students. Furthermore, students who practiced calibration should be more accurate by topic and that students who only calibrated their scores on the final SOL test would show more variability in their accuracy by topic. For example, students who practiced calibration would have more stable prediction accuracy across topics than those that did not practice calibration. It was also hypothesized that student self-efficacy beliefs at the beginning of the course would predict calibration bias, but not performance. Finally, it was anticipated that students would find the topical calibration practice helpful in increasing their accuracy and performance.

### Summary and Contributions of Present Study

This chapter included a review of types of calibration measurements and the ways these measurements are calculated. It has also examined the correlation between achievement and self-efficacy, and the effects of strategies or interventions aimed toward increasing student calibration accuracy and test performance. Calibration is a metacognitive function that is closely linked with self-efficacy and performance outcomes. Klassen (2006) refers to calibration as "...the congruence of self-efficacy beliefs with subsequent performance" (p. 183), and the literature supports this definition. Stolp and Zabucky (2009) suggest that the construct of self-efficacy may be all but impossible to separate from student calibration of comprehension. Both self-efficacy (Pajares, 2006), and calibration (Chen, 2003) have been shown to explain a large percentage of variance in academic achievement. It seems logical that if strategies are implemented that improve both self-efficacy and calibration accuracy then test performance should be improved as well.

Past studies have had mixed results as to whether calibration practice increases calibration accuracy (Bol et al. 2010; Hacker et al. 2008). In 2000, Hacker et al. suggested that not enough focus had been given to students' prediction accuracy and to improvements in accuracy when students are asked to make judgments over longer periods of time under more motivating circumstances. Eight years later, Hacker et al.(2008) modify this somewhat by suggesting that "Simply providing with students practice tests and feedback on calibration accuracy is not enough to significantly improve their accuracy" (p. 450). The authors suggest that explicit training in monitoring may be

needed. The current research attempted to train students in explicit monitoring by asking students to calibrate their test scores on each test topic several times over the course of a semester. As they recorded their topical predictions they were encouraged to measure them against previous predictions and actual results. Practice culminated in the calibration of a high-stakes math test, which should have been very motivating to most students.

In addition, this researcher sought to determine if increases in calibration accuracy and achievement on high-stakes tests were possible for culturally diverse, urban, adolescent students who may be performing poorly. Other studies have addressed college level populations, or less diverse suburban populations. Students in college settings are obviously more motivated to perform well academically, or else they would likely not be in college. Fewer studies have been conducted with urban, lower achieving student populations, and this group of students may be the most in need of metacognitive strategies to increase their self-efficacy and calibration accuracy.

The methodology for investigating the research questions follows in Chapter III. Chapter III provides specific information on the participants, measures, procedures, and analyses that were used in the study.

## CHAPTER III

### METHODOLOGY

#### Introduction

Whether practice with a topical calibration strategy would improve middle school adolescent students' calibration accuracy or their performance on an end-of-course high stakes math test were the primary questions addressed. In addition, this study examined the relationship between calibration accuracy and achievement. Another aspect of the research was to determine if student self-efficacy predicts student performance or calibration accuracy. Also addressed was whether students who practiced calibration would do better predicting their scores for each topic than students who did not practice calibration, or if students who practiced calibration were better calibrated for some math topics than others. More specifically, the following research questions were addressed:

1. Does topical calibration practice affect student calibration accuracy or test performance?
2. Do topical calibration practice and achievement level interact to influence prediction accuracy?
3. Does calibration accuracy differ by topic and does accuracy by topic differ by treatment condition?
4. Does self-efficacy of students who practiced by topic predict calibration accuracy and SOL performance?

5. What are the student's beliefs about the efficacy of calibration practice on their accuracy and performance on tests?

This chapter will delineate the participants, design, measures, and procedures that were used to answer these research questions. It begins with a description of the participants followed by the design, measures, and procedure for the study.

#### Schools and Participants

A total of 110 middle school students enrolled in regular sixth grade math at a Norfolk public middle school participated in the study. According to the state's enrollment demographics for the 2010-2011 school years, the majority of the 796 students at this school identified themselves as Black (43%), followed by White (33%), then Hispanic (13%), with the remaining 11% divided between being of two races or more, Asian, American Indian, or Native Hawaiian/Pacific Islander. Of the 796 students, 404 are male and 392 are female.

Three sixth grade regular math teachers participated in the study. One of the teacher's classes was randomly assigned to the practice condition ( $n = 53$ ) where the students had the opportunity to practice topical calibration on three tests plus the SOL test. The other teachers' classes ( $n=57$ ) only calibrated their scores topically once on the final SOL test.

Although students were tracked by name across the study, all identifying information was stripped from the data once it had been compiled in order to protect the confidentiality of the students.



## Design

This was a quasi-experimental, correlational study conducted over the course of three months. It was quasi-experimental because an intervention was introduced in the form of a calibration strategy. The study used a topical calibration strategy requiring students to predict their performance on each topic of the test. For example, students predicted how many questions they would answer correctly for each topic (i.e. statistics, when given the number of questions in that topic). Students then added their predicted scores for each topic on the test for a global test prediction.

It was also correlational because it was expected that the data would support relationships between variables such as self-efficacy and calibration accuracy. The independent variables were whether students calibrated by topic, self-efficacy, and achievement level (low performers vs. high performers). The dependent variables were calibration accuracy and bias, and the performance of students on the high stakes Math SOL test. Students responded to a 5 question self-efficacy inventory at the beginning of the study in addition to calibrating their scores for tests.

One teacher (53 students) participated in the calibration practice condition across three tests and then the final SOL and two teachers (57 students) only had students calibrate their scores on the final SOL. This was to help isolate the effect of calibration practice on calibration accuracy. Although it was possible that there were teacher effects arising from having different teachers for the two conditions (calibration practice vs. no practice), it was not expected to be a major limitation. The subject matter and tests were standardized, resulting in teachers having to closely follow the curriculum to ensure

students were prepared for tests. Therefore, students were taught the same material, at the same time, and in the same way.

## Measures

### *Self-Efficacy Measurement*

Pajares and Miller suggest that the self-efficacy assessment may in itself be useful if calibration and self-efficacy work in concert to influence how students consider their metacognitive capabilities (1997). To determine if self-efficacy and calibration accuracy are correlated with performance, self-efficacy assessments were administered to all participating students.

The Academic Efficacy Scale developed by Midgley, et al. (2000), as part of the Patterns of Adaptive Learning Scales (PALS), was used for this study (Appendix A). The Academic Efficacy Scale was distributed to all of the participating students prior to the beginning of the study. An example of a self-efficacy question is, “I’m certain I can master the skills taught in class this semester.” Students were asked to rate statements on a 5-point Likert-type scale ranging from “not at all true” (1) to “very true” (5).

There are 5 questions in total. The Cronbach alpha measure of internal consistency reliability was .78 for the scale. The authors measured it against other self-efficacy scales for construct validity and found it consistent with other research (Midgley, Kaplan, Middleton, Maehr, Urdan, Anderman, Anderman, & Roeser, 1998). This scale will be adapted by focusing the questions specifically for math. For example, the original

question above was restated to read “I’m certain I can master the skills taught in Algebra this semester.”

### *Calibration Materials and Measures*

Students in the topical practice calibration condition received a calibration worksheet prior to their 3rd quarterly test, one unit test, and a Mock SOL test, to prepare them for their final calibration of the end-of-course SOL test. The worksheet listed the test topics next to a box for students to write how many questions they expected to answer correctly for each topic. Topical calibration worksheets used for the practice and no practice conditions are available in Appendix A.

### *Math Achievement*

The sixth grade students took math SOL exam at the end of the year. High school students in Norfolk Public Schools are required to take and pass three math courses for a standard graduation diploma, and four math classes for an advanced diploma, at the level of algebra or higher. In addition all NPS students must earn a minimum of one verified credit for standard diplomas, and two for advanced diplomas, in a math class at the algebra level or higher. A verified credit is earned when the student passes the corresponding SOL test for a given course. The sixth grade math class is used as a placement tool for students for the next year’s math class. If students do well enough on the sixth grade test they are promoted to Algebra in seventh grade.

SOLs are criterion-based tests to assess whether students meet specific minimum expectations on state standardized objectives. In 2009, the State of Virginia compared the State’s SOL tests to the Common Core State Standards (CCSS) for math, for all grades,

and found that the two sets of standards are aligned. They also indicated that Virginia's standards are, in some cases, more rigorous than those of the CCSS, helping to ensure the validity of the tests.

Tests are developed by committees of reviewers, who are nominated and representative of educators throughout the state. The reviewers recommend test items. The test questions are field tested yearly by the Virginia State Department of Education (DoE) to enhance the reliability and validity of the assessments. The SOLs are aligned with the objectives found in the Virginia SOL Blueprint document (Virginia Department of Education, 2002). The math SOL tests usually consist of 45-50 multiple-choice questions, with 10 additional field test questions that are not counted in the score.

Students additionally take unit tests, called Common Formative Assessments (CFAs) developed by the math teams at each school for each grade and subject. These tests are then forwarded to the Department Chair in math, who reviews, modifies, and approves them as appropriate. These assessments are based on the SOL objectives listed in the SOL blueprint documents for the specific unit students are studying at that time. Students participating in the practice condition took one CFA test prior to the SOL. In addition, students took one quarterly test at the end of the third quarter. This test is also based on SOL objectives, and is a cumulative test of the concepts students learned during the quarter. Quarterly tests are developed by staff at the district level.

### *Student Perceptions of Calibration Efficacy*

After completing their prediction forms for the SOL tests at the end of the semester, students in the practice calibration condition were asked their perceptions of the efficacy of their specific calibration practice. The three questions were:

1. Practice in predicting my overall test scores helped me think about how to do better on the next test. (Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

2. Practice in predicting how well I would do helped me think about what areas I needed to work on to do better on the next test. (Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

3. Practice predicting my test scores helped me to do better on my tests. (Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

### **Procedure**

The researcher met with the math department chair at the school at the start of the study (March 2011) to explain the study, review the materials (calibration worksheets, self-efficacy scales, parent opt-out forms), and answer any questions.

Prior to beginning the study, students were given a "Parent Opt-Out" Form (Appendix B). This is a standard form that the school division uses for research studies to provide parents with the option to remove their children from the study. If the opt-out

form was not returned by the end of the second class period, the students were expected to participate in the study. No students opted out of the study.

All participating students in the practice condition completed the Self-Efficacy scale (Appendix C) prior to the study. The teacher read a script (Appendix D) verbatim prior to distributing the self-efficacy scale at the beginning of the study. The script was developed by Midgley et al. (2000) and adapted for the purposes of this study by referencing math classes in particular. The script advised students that the survey was not a test and that there were no right or wrong answers; students were reassured that the information that was collected would remain confidential and that their parents and peers would not see their specific responses to any of the questions. Students were told that some questions may have sounded very similar to others in the survey, but that this was important for ensuring a good understanding of what each student thinks. A sample question was included and reviewed with the students to familiarize them with the Likert scale.

Participating students started predicting their test performance beginning with the first quarterly test after the deadline for opt-out forms had passed. Calibration forms were passed out just prior to the distribution of the quarterly test, as well as the CFA Unit test and the Mock SOL. Students began calibrating their scores on the quarterly test for practice. The instructions were given at the beginning of the study and then again before each calibration event. The teacher explained the calibration procedures initially (see Appendix E), and again for the first testing event, but for all remaining tests, the

calibration worksheets were passed out and if students were absent previously or had questions, they were instructed to ask their teacher.

Students were provided with a list of topics on the exam, and examples of types of problems for each topic. They were asked to predict their scores (number of items they expected to answer correctly) for each topic. The number of items was listed for each topic. When the tests were scored the department head entered the number of problems answered correctly for each topic on student calibration worksheets. The worksheets were returned to the students when the tests were reviewed, and students were encouraged to compare their predictions with their actual scores for each test.

Students in both the practice and no-practice conditions were provided with a topical prediction worksheet for the high-stakes final SOL test. Only a few students had the opportunity to get their EOC SOL scores back in time before the end of the school year, so the researcher had to enter most of the actual scores from the test following the study. This form is available in Appendix F.

In the class immediately preceding the SOL test in May, students were also asked three questions and responded using a Likert Scale with four possible answers ranging from Strongly Agree to Strongly Disagree. The items assessed their perceptions of the effectiveness of the treatment in increasing their achievement and causing them to consider potential areas of weakness prior to the test. These questions were listed previously, in the instrumentation section, but are also available in Appendix G.

Students were asked to put their name on all of the materials, as well as their teacher's name. Once the data was collected, the names were replaced with identification numbers protect participant confidentiality.



## CHAPTER IV

### RESULTS

#### Introduction

The results of the analyses used to evaluate the effectiveness of topical calibration on the calibration accuracy and achievement of middle school students on tests are presented in this chapter. Self-efficacy was additionally analyzed to determine if it predicts either calibration accuracy or achievement. A total of 110 sixth grade regular math students participated in topically calibrating their SOL scores at the end of the course, and 52 of those only calibrated their test scores prior to the SOL. The 52 students that practiced topical calibration did so for three tests prior to the SOL: a quarterly exam, a unit exam, and a Mock SOL exam. The students that practiced topical calibration prior to the SOL are referred to as the practice condition. The students that only calibrated their scores for the SOL at the end of the course are termed the no practice condition. There were five topics on the SOL test: Numbers and Number Sense; Computation and Estimation; Measurement and Geometry; Probability and Statistics; and Patterns, Functions, and Algebra.

The analyses began with the impact of the practice condition on achievement. Descriptive statistics are presented on the test scores for both the practice and non-practice conditions. The descriptive statistics are followed by the results of an analysis of variance (ANOVA) for overall achievement and then a multiple analysis of variance

(MANOVA) for achievement on each of the topic areas. Next, the results of the impact of practice condition on calibration accuracy are described, beginning with descriptive statistics for the absolute differences and signed directions (bias) in prediction accuracy. To analyze the impact of treatment on accuracy, another ANOVA was conducted using total scores, followed by a MANOVA for topical calibration scores. A median split was used to divide the students into two groups, those who score above the median on the SOL test (high achievers) and those that scored below (low achievers). Using this categorization, an ANOVA was conducted. The ANOVA analyzed the impact of achievement level and practice condition on calibration accuracy for total SOL test score predictions. This was followed by a MANOVA for calibration accuracy by topic, again using high or low achievement on the test as one independent variable and practice condition as the other independent variable. Following this, two regressions were performed to determine if predictive relationships exist between self-efficacy and achievement or self-efficacy and calibration accuracy. The chapter concludes with descriptive statistics that reflect student perceptions of the topical calibration practice strategy.

### The Impact of Topical Calibration Practice on Achievement

#### *Descriptive Achievement Results*

To begin the analysis of achievement results that compare the practice conditions (calibration practice and no-practice); the means for both groups were calculated. The results for SOL test (the end-of-course Standards of Learning test) are displayed in Table 1. Recall that students were asked to predict how many items they expected to answer

correctly for each topic on the test; therefore the means represent the number of problems that students correctly answered out of 50 problems on the test. The means between the practice condition and the no practice condition were very similar: the group that practiced calibration had a mean of 31.4, and the mean for the group that only predicted their scores on the SOL test was 31.6, with standard deviations of 9.7 and 9 respectively. Considering that a passing score rating of 'Proficient' for this test required that students correctly answer 33 problems, the mean scores for both groups are below the score needed to pass.

Table 1

*Average SOL Scores for Calibration Practice and No Practice Groups*

Condition	<i>n</i>	Mean	SD
Practice Calibration Group	53	31.4	9.7
No Practice Calibration Group	57	31.6	9.0

Although students in the practice calibration group did not do well on the final SOL test, they did slightly better than they did on prior tests. They practiced topical calibration on three tests throughout the semester prior to the SOL test. Calibration on these tests was the experimental manipulation. The first test was a quarterly test with 50 items, the second test was a unit test with 20 items, and the third test, just prior to the SOL was a mock SOL practice test with 50 items. They were asked to predict their scores for each subject on the tests prior to taking each of the three tests. Students did not score very well on any of these tests: the mean for the first test, a quarterly test was 21 with a standard deviation of 10.6. The mean for the unit test was 14.7 with a standard deviation of 3.9. The mean for the third test was 27.4 with a standard deviation of 10.2. The Unit

Test was a shorter test than the other tests, and student achievement was better for that test than the others, with a mean of 14.7 out of 20 items (73.5%). This can be compared to the mean Quarterly Test score of 43.4% or the mean Mock SOL score which was 54.8%.

#### *Achievement by Treatment Condition*

As foreshadowed by the descriptive statistics, the ANOVA results for achievement did not show a significant difference between the practice condition and the no practice condition for test scores. A MANOVA was performed to determine if there was a significant difference on the numbers of problems answered correctly for each topic on the SOL test when comparing conditions. There were no significant differences on any of the SOL topics for practice versus no practice groups. The results of these analyses are shown in Table 2.

Table 2

*ANOVA and MANOVA Results for Calibration Practice and No Practice Groups on SOL Test*

	df	F	Sig.
<u>ANOVA</u>			
Global SOL Test Achievement	1,109	.008	.93
<u>MANOVA</u>			
<u>Test Achievement by Topic</u>			
Numbers & Number Sense	1, 109	1.9	.18
Computation and Estimation	1,109	.00	.96
Measurement and Geometry	1,109	.37	.55
Probability and Statistics	1,109	.01	.93
Patterns and Functions	1,109	.00	.98

## The Impact of Topical Calibration Practice on Calibration Accuracy

### *Descriptive Calibration Accuracy Results*

Descriptive statistics consisting of mean absolute accuracy and mean signed accuracy with their standard deviations are shown in Table 3. Recall that absolute differences show only the difference between the prediction and actual score without respect to the direction of the score (overconfidence or underconfidence). The signed differences show the direction (bias) of the difference between the prediction and actual score. For example, if the student predicted they would get 9 problems right, and they only got 6 right, the difference is 3 ( $9-6$ ). A positive score, such as in the example just given, reflects overconfidence; whereas, a negative score (for example if the student had gotten 10 right instead of 6, they would have had a result of  $9-10 = -1$ ) reflects underconfidence. The global accuracy of both the practice and no practice conditions was very close; the practice condition was within 9 points of their predicted scores overall while the no practice condition calibrated slightly better, within 7 points of their predicted scores. Both conditions were overconfident, although the practice condition showed more overconfidence with a signed accuracy mean of 7.3 compared to the no practice condition signed mean of 4.3.

Table 3

*Descriptive Statistics for Calibration Accuracy and Direction between Calibration Practice and No Practice Groups on SOL Test*

	Absolute Accuracy			Signed Accuracy	
Condition	<i>n</i>	Mean	SD	Mean	SD
<u>Total SOL Test Accuracy</u>					
Calibration Practice Group (CP)	52	9.5	8.7	7.3	10.7
No Practice Group (NP)	57	7.4	6.8	4.3	9.4
<u>SOL Topical Test Accuracy</u>					
Numbers & Nmbr Sense CP	51	1.8	1.8	1.2	2.3
Numbers & Nmbr Sense NP	57	1.7	1.3	.3	2.1
Computation & Estimation CP	51	2.3	1.9	1.2	2.7
Computation & Estimation NP	57	2.4	2.0	1.0	3.0
Measurement & Geometry CP	51	3.0	2.2	1.8	3.3
Measurement & Geometry NP	57	2.6	1.9	1.4	2.8
Probability & Statistics CP	51	1.9	1.6	1.6	1.9
Probability & Statistics NP	57	2.1	1.6	.7	2.6
Patterns & Functions CP	51	2.5	2.3	1.4	3.1
Patterns & Functions NP	57	2.8	2.3	.8	3.5



Results of descriptive statistics by topic were similar to the total means in that means did not differ much between the practice and no practice conditions, and both groups were slightly overconfident across topics. Students in the calibration practice condition were marginally more accurate on three of the five topics than students in the no practice condition, but students in the no practice condition were less overconfident for each topic. Students were best calibrated in the Numbers and Number Sense Topic. The no practice condition was slightly more accurate with a mean of 1.7 than the practice condition's mean of 1.8, and they were less overconfident at .3 than the practice condition at 1.3. Students displayed the most inaccuracy on the third SOL topic, Measurement and Geometry. The no practice condition was again slightly more accurate with a mean difference score of 2.6 compared to the practice condition's mean score of 3, and the no practice condition was less overconfident again with a mean signed difference score of 1.4 compared to the practice condition's signed difference score of 1.8.

#### *Calibration Accuracy by Treatment Condition*

To determine if the differences between the practice and no practice conditions were significant for absolute accuracy on total SOL scores an ANOVA was performed. The result of the ANOVA was not significant,  $F(1,109) = 1.62, p > .05$ . A MANOVA was performed to test for significant differences between the conditions for each topic on the SOL. No significance was found by condition for any of the topics. The results of this ANOVA and MANOVA are shown in Table 4.

Table 4

*ANOVA and MANOVA Results for SOL Calibration Accuracy*

	df	F	Sig.
<u>ANOVA</u>			
Global SOL Test Accuracy	1,108	1.96	.16
<u>MANOVA</u>			
Numbers & Number Sense	1, 107	.34	.56
Computation & Estimation	1,107	.11	.74
Measurement & Geometry	1,107	.1.35	.25
Probability & Statistics	1,107	.77	.38
Patterns & Functions	1,107	.52	.47

**The Impact of Achievement Level on Calibration Accuracy**

To examine the relationship between achievement level and absolute calibration accuracy, students were split into two groups based on the median score (31.5 out of 50 problems correct) and characterized as either higher or lower achievers accordingly.

There were 55 high achievers (scoring 32 or above) and 55 low achievers (scoring less than 32). One ANOVA was conducted to analyze the effects of the independent variables of practice condition and achievement level (high or low) on prediction accuracy for total SOL scores.

The ANOVA revealed a significant main effect for achievement level on absolute calibration accuracy,  $F(1,108) = 35.07$   $p > .00$ . High achievers in both conditions were significantly more accurate than low achievers across conditions. There was no main effect for condition. However, the interaction between achievement level and condition was statistically significant. The results are shown in Table 5.

Table 5

*Effect of Achievement (high or low) and Practice Condition on SOL Total Calibration Accuracy*

	df	F	Sig.
Achievement Level	1,108	35.07	.00
Practice Condition	1,108	3.27	.073
Condition X Achievement	1,108	4.41	.038

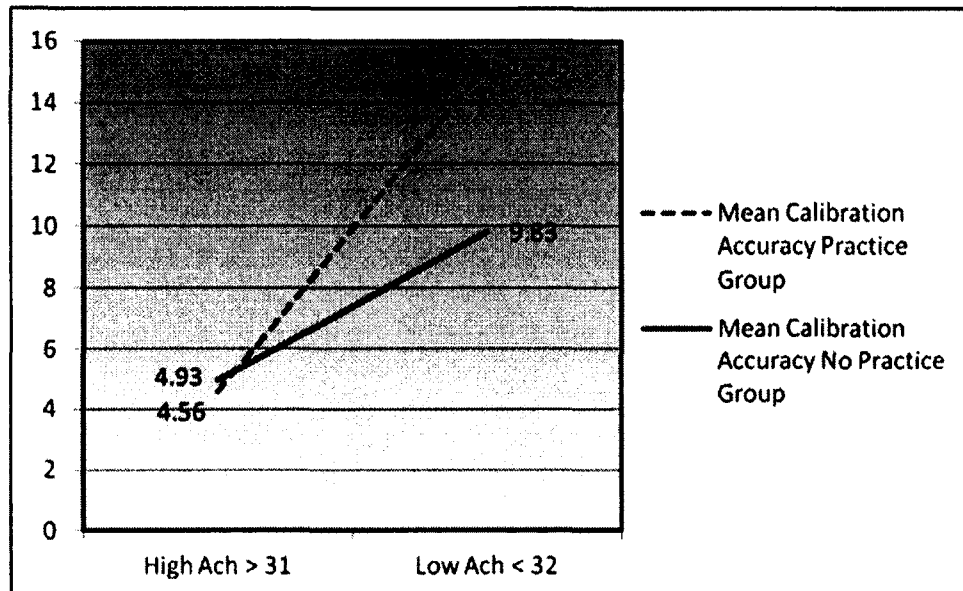
Descriptive statistics to examine the group means based on achievement and condition are shown in Table 6. The results are also graphically depicted in Figure 1. There was almost no difference in absolute accuracy between high achievers in either practice condition; they were both fairly accurate at within 5 points of their predicted scores. There was a difference in absolute accuracy between low achievers in the practice and no practice conditions, with participants in the no practice condition being more accurate. Both low achieving groups were much less accurate than the high achievers with the practice condition 15 points on average from their predicted scores and the no practice condition 10 points on average from their predicted scores. Both groups were overconfident in their predictions, with the low achievers showing more over confidence.

Table 6

*Total Calibration Accuracy Means and Standard Deviations for High/Low Achievement Groups*

Condition	High Achievers (Score > 31)			Low Achievers (Score < 32)		
	<i>n</i>	Mean	SD	<i>n</i>	Mean	SD
Practice Group	27	4.56	4.13	26	15.23	9.24
No Practice Group	28	4.93	4.26	29	9.83	7.83

Figure 1 shows absolute accuracy means by practice condition and achievement level graphically.



*Figure 1.* Absolute Calibration Accuracy by Achievement Level and Practice Condition

A MANOVA was conducted to see if student calibration accuracy differed by condition or achievement level on any of the 5 topics. The results for the MANOVA are shown in Table 7. The difference in calibration accuracy was significant for each topic between the higher and lower achievers. The difference between the practice and no practice conditions was not statistically significant for any topic.

Table 7

*Multiple Analyses Of Variance showing Effect of Achievement Level (high or low) and Practice Condition on SOL Topical Calibration Accuracy*

	df	F	Sig.
<u>Achievement Level (high/low)</u>			
Numbers & Number Sense	1, 107	19.89	.00
Computation & Estimation	1,107	31.83	.00
Measurement & Geometry	1,107	12.67	.00
Probability & Statistics	1,107	23.18	.00
Patterns & Functions	1,107	12.22	.00
<u>Practice Condition (P/NP)</u>			
Numbers & Number Sense	1,107	.51	.48
Computation & Estimation	1,107	.08	.78
Measurement & Geometry	1,107	1.65	.20
Probability & Statistics	1,107	.80	.37
Patterns & Functions	1,107	.48	.49

The means and standard deviations for absolute calibration accuracy on each topic of the test for high and low are displayed in Table 8. The high achievers were usually closer by more than one point and sometimes two points from their predicted scores for

each topic on the test when compared to the low achievers. However, both groups were relatively accurate but again overconfident, with high achievers slightly less so.

Table 8

*Topical Calibration Accuracy Descriptive Statistics: High and Low Achievement Groups*

SOL Test Topic	Higher Achievers (Score > 31)			Lower Achievers (Score < 32)		
	<i>n</i>	Mean	SD	<i>n</i>	Mean	SD
Numbers & Number Sense	54	1.13	1.0	54	2.33	1.8
Computation & Estimation	54	1.43	1.2	54	3.30	2.1
Measurement & Geometry	54	2.11	1.4	54	3.44	1.7
Probability & Statistics	54	1.35	1.3	54	2.67	2.5
Patterns & Functions	54	1.91	1.9	54	3.37	1.8



## The Relationship between Self-Efficacy, Achievement, and Calibration Accuracy

### *Descriptive Self-Efficacy Results*

Students in the topical calibration practice condition completed a self-efficacy scale prior to their first calibration practice attempt. Fifty-one students completed the scale. There were five statements on the scale that required students to choose a point on a number scale from 1 to 5 that reflected their confidence for that statement. For example, the first statement was “I am certain I can master the math skills taught in class this semester”. Students chose an answer from 1 (Not At All True) to 3 (Somewhat True) to 5 (Very True). The mean self-efficacy score was 3.5, with a standard deviation of .92. Students were slightly more positive than negative about their math self-efficacy. Descriptive statistics for the questions on the scale are presented in Table 9.

Table 9

*Descriptive Statistics for Self Efficacy Scale Question Answers (n =51)*

Self Efficacy Scale Question	Mean	SD
I am certain I can master the math skills taught in class this semester	3.34	1.1
I am certain I can figure out how to do the most difficult math class work	2.98	1.1
I can do almost all the work in math class if I don't give up	3.85	1.3
Even if the work in math class is hard, I can learn it	3.51	1.3
I can do even the hardest work in math class if I try	3.80	1.3
Overall	3.50	.92

Two regression analyses were conducted to determine if self-efficacy predicted student achievement or calibration accuracy. The average self-efficacy score was entered into the regression model with actual SOL scores for the treatment group as the criterion variable. Self-efficacy did not significantly predict achievement for this group,  $\beta = .24$ ,  $t(1.55) = .130$ ,  $p > .001$ . The results of the second regression model using self-efficacy as the predictor variable for absolute calibration accuracy as the criterion variable was

also not statistically significant,  $\beta = .33$ ,  $t(-.204) = .840$ ,  $p > .001$ . Student self-efficacy beliefs did not predict achievement or prediction accuracy for the SOL Test. The regression results are in Table 10.

Table 10

*Regression: Self-Efficacy Belief as a Predictor of Achievement or Calibration Accuracy*

Condition	df	F	Sig	R <sup>2</sup>
Achievement	1,39	2.39	.130	.06
Calibration Accuracy	1,39	.042	.840	.00

#### Student Perceptions of Treatment Effectiveness

Students in the treatment condition were asked three questions about the effectiveness of the treatment, and asked to choose whether they strongly disagreed (1), disagreed (2), agreed (3), or strongly agreed (4) with item. The mean overall response was a 2.83, indicating more agreement than disagreement with the questions. The items follow.

1. Practice predicting my overall test scores helped me think about how to do better on the next test.
2. Practice in predicting how well I would do helped me think about what areas I needed to work on to do better on the next test.
3. Practice predicting my test scores helped me to do better on my tests.

The descriptive statistics are shown in Table 11. The mean values obtained across the 3 items were very close. Students wavered between disagreeing and agreeing, although there was more overall agreement with questions one and two than three. Collapsing the agreement categories shows that 75% of students agreed that prediction practice helped them reflect on how to do better on the next test. A total of 83% of students agreed that prediction practice helped them think about what areas they needed to work on to do better on the next test. In contrast, only just over half, 54% of the students agreed that prediction practice helped them do better on their tests.

Table 11

*Descriptive Statistics for Student Perceptions of Treatment Effectiveness Answers*

	Item 1 <i>n</i> = 47	Item 2 <i>n</i> = 46	Item 3 <i>n</i> = 46
Mean Score	2.94	2.96	2.59
Strongly Disagree	8%	6%	13%
Disagree	17%	11%	33%
Agree	47%	63%	37%
Strongly Agree	28%	20%	17%

## CHAPTER V

### DISCUSSION

#### Overview

This study was an attempt to determine whether urban middle school math students improve their performance and calibration accuracy on tests with topical calibration practice and determine whether a relationship exists between self-efficacy and performance or calibration accuracy. This chapter begins with a discussion of how calibration practice influenced test performance. The effects of calibration practice on calibration accuracy will be discussed next, followed by a discussion about achievement level and calibration accuracy. This will be followed by a discussion of self-efficacy as a predictor of achievement or calibration accuracy. Student perceptions of the effectiveness of the topical calibration practice are discussed next. The chapter concludes with a discussion of the limitations of the study along with directions for future research.

#### Calibration Practice and Test Performance

Nietfeld et al. (2006) asserted that “students who improved their calibration also had a strong tendency to improve their performance on class tests” (p172). However, this study’s hypothesis that students who practice calibration would perform better on their SOL test than students who only made predictions on the final SOL test was not supported. Test performance on the high-stakes Standards of Learning Test at the end of the semester was very similar between the calibration practice condition and no practice condition. The SOL test has 50 problems. Students in the calibration practice condition

had a mean test score of 31.4 out of the 50 problems correct, and those in the no practice condition had a mean test score of 31.6. Both of these means are below 33, the score needed to pass the test. It was hoped that the introduction of a topical calibration strategy would improve student test performance in the practice condition, but this was not the case. This was not completely unexpected, as other research findings have been inconclusive as to whether calibration strategies enhance test performance. Bol et al.'s study (2005) on the effects of overt calibration practice on college students' accuracy and test performance did not yield significant results for either accuracy or performance. The mean percentage scores on final examinations between the calibration practice (covert) and no practice groups (overt) were virtually identical. Huff and Nietfeld's study (2009) with fifth grade students also failed to find a difference in performance between the control and treatment groups. The treatment groups used calibration strategies that included during-task monitoring and after-task monitoring. Although all groups did improve their performance from the pre to the post test, there were no significant differences between the performance of the treatment and no treatment groups.

However, some studies have indicated that an improvement in performance may be linked to calibration strategies. In Bol et al.'s calibration study (2012) with high school biology students, half of the students were provided with calibration guidelines that asked students to consider their understanding of the material as well as their strengths and weaknesses related to their content mastery. A second condition was whether students practiced calibration in individual or group settings. Thus, students were assigned to four groups: 1) students that used guidelines in group settings, 2) students that

used guidelines individually, 3) students that worked in groups without guidelines, or 4) students that worked individually without guidelines.

The results largely confirmed the authors' hypothesis. Students that calibrated in groups consistently scored higher on the tests than students who calibrated individually. It may be the case that placing students in groups may encourage students to seek help from peers (Bol et al., 2012). Calibrating scores in a group setting also allows students to get peer feedback and may help them evaluate their own level of knowledge compared to others in their group. With respect to the present findings, topical calibration practice may not have been enough. Perhaps a topical calibration strategy would be more efficacious if it was used in a group setting. That would allow students to collaboratively discuss their strengths and weaknesses on the test. On the other hand, there are some potential drawbacks to group work. Puncochar and Fox's (2004) research indicated that although groups were more accurate in their calibration, they were more overconfident about their wrong answers, and their overconfidence increased across quizzes.

Bol et al.'s (2012) study further showed that calibration guidelines improved test performance. It was hoped that the topical strategy used in the current study would encourage the same type of metacognitive deliberation as Bol's guidelines seemed to have done for students. However, though it was expected that prediction of scores on topics would promote deliberation, the cues may not have been salient enough. A more straightforward prompt such as calibration guidelines might be more beneficial in urban school settings. The guidelines asked the students directly to consider their strengths and weaknesses about the material, whereas the topical calibration strategy required they



predict their score for each topic on the test. They were not asked to synthesize that knowledge and determine areas of strengths and weaknesses. It is possible that students did not take the time to reflect on their topical predictions, rendering their usefulness for increasing achievement negligible.

Nietfield et al.'s study (2006) with monitoring exercises had similar results to Bol et al.'s (2012), indicating that reflection can be effective if more in-depth. College students in the treatment condition were provided with monitoring exercise worksheets. The worksheets asked students to rate their understanding of the day's content, identify concepts they found difficult, explain what they would do to improve their understanding of these concepts, and answer three multiple-choice review questions with confidence judgments for each. The students' scores on the first test were very close across groups, but on the second test, students in the treatment group scored one standard deviation higher than the comparison group and maintained the difference through the third and fourth tests.

Hacker et al. (2000) suggest another reason for the lack of improvement in performance. Students do not use the results of their monitoring to regulate their future test preparatory activities. If this is true of a college age population, than it probably is true of a younger population that has even less experience with self-regulation, like performance monitoring. A possible solution is to require students to graph their predictions and results over several tests. This strategy is similar to the one employed by Brookhart et al. (2004). These researchers had two classes of third grade students graph both their predictions and results (using bar charts) on multiplication tests over 10 weeks.

The overall average test scores rose in both classes. Similarly, Zimmerman (2006) hypothesized that “graphing can enhance a learner’s sensitivity to improvements in functioning” (p.211), resulting in gains in performance. This was supported by his results. Participants who graphed their predicted scores and actual scores performed better than students who did not participate in the graphing. Perhaps the combination of more practice and graphing would result in more careful self-monitoring and regulation of test taking preparation or behavior.

#### Calibration Practice and Calibration Accuracy

Whether calibration accuracy can be substantially improved is still undetermined. Hacker et al. (2009) suggest that the difference between classroom-based studies that show improvement in calibration accuracy versus those that do not show improvement lies in the power or strength of the intervention (p. 446). For example, they suggest that reflection and instruction on self-assessment and monitoring strategies were effective in some studies (Nietfeld et al., 2006) at improving accuracy, but not in others (Hacker et al., 2000, Bol et al., 2010). The topical calibration practice strategy used in this study did not improve students’ calibration accuracy for the SOL test. Though the difference was not statistically significant, students in the no practice condition were more accurate in their predictions for the SOL than students in the practice condition. Students who practiced topical calibration on three tests prior to the SOL had a total mean accuracy score of 9.5 on the SOL test, whereas, students who didn’t practice calibration had a mean accuracy score of 7.4, almost a 2 point difference.

One of the hypotheses of the present study was that calibration practice would improve students' calibration accuracy on each topic of the test, but this prediction was not supported. Calibration practice did not improve accuracy at the topical level; both conditions were within half a point of each other for each topic. For example, the practice condition had a mean accuracy score of 2.3 for the topic Computations and Estimations, while the no practice condition had a mean accuracy score of 2.4.

While frustrating, the resistance of calibration accuracy to improvement has been documented by other researchers (Bol et al., 2001; Bol et al., 2005; and Bol et al., 2010). It is difficult to ascertain why calibration practice does not always result in improved accuracy, but several reasons have been promulgated. Hacker et al. (2000) hypothesized that students would base their performance expectations on prior judgments of performance rather than actual prior performance. The researchers expected this to change as the semester progressed and students obtained actual performance experience, but it did not. Students continued to make performance judgments based on their prior judgments of performance rather than their actual performance, even though higher performing students did show modest gains in prediction accuracy. Bol et al (2005) echoed this sentiment, suggesting that instead of basing performance judgments on objective feedback, students may base predictive judgments on persistent feelings of their own learning attributes (p. 270). Evidence for this was also provided anecdotally in the present study when one student commented that practice at predicting their score didn't make a difference because they continued to rely on a benchmark score of 90, regardless of test outcome. This is characteristic of the literature on explanatory or attributional

style. An attributional style is an individual's explanation of outcomes such as success or failure, i.e. "I didn't pass the test because it was tricky". Hacker et al.'s study (2008) investigated the effect that explanatory style has on calibration accuracy, along with reflection strategies and extrinsic rewards. They found that when all students were considered as a whole, none of their interventions were effective at increasing calibration accuracy. However, when students were separated into groups based on performance, lower performing students' calibration accuracy increased when extrinsic rewards were offered. They also found correlational evidence that attributional style (e.g. concerns over studying behaviors and social influences) explained a significant amount of variance in prediction and postdiction accuracy.

Another reason for the lack of improvement in calibration accuracy, specific to the present study, was that students were taking multiple tests over the course of two or three weeks, including other SOLs and year-end testing. This may have resulted in students expending less effort on formulating accurate predictions specific to this SOL and relying on prior judgments of performance instead, particularly if they felt overwhelmed by tests. This idea is supported by Hacker et al. (2000) who posited that accuracy may be reduced when people are faced with complex memory demands.

Overconfidence is a common and well-documented problem, especially amongst lower achievers (Bol et al., 2005; Chen, 2003; Grimes, 2002; Hacker et al., 2000; Klassen, 2006; Kruger & Dunning, 1999). It was hoped that a topical calibration strategy would encourage more reflective thinking about strengths and weaknesses for specific topics on the tests, and result in more accuracy and less overconfidence. Yet this was not

the outcome. All of the students expressed overconfidence in their total and topical prediction similar to results reported in Bol et al., 2010; Chen, 2003; and Klassen, 2006. Surprisingly, students in the no practice condition expressed less overconfidence; the practice condition had a signed mean accuracy score of 7.3, while the no practice condition had a signed mean accuracy score of 4.3. This finding is reminiscent of those found in Huff and Nietfeld's (2009) study with fifth graders. The researchers were surprised when one of their treatment groups "...showed a significant increase in bias towards overconfidence" in their postdiction calibration (p. 172). They suggested that the treatment may have led to student overconfidence that was not commensurate with their ability. It is possible that students participating in the calibration practice group also believed that calibration practice would automatically increase their test scores without adequately considering whether their ability had improved.

#### Achievement Level and Calibration Accuracy

As hypothesized, a significant interaction was found between condition and achievement. Students were split into two groups based on the median SOL Score (31.5 out of 50 problems correct) and characterized as either higher or lower achievers accordingly. There were 55 high achievers (scoring 32 or above) and 55 low achievers (scoring less than 32). High achievers in both conditions (Practice and No Practice) had a mean accuracy score less than 5 points from their actual scores. Low achievers were much less accurate. In the practice condition the average accuracy score was 15.23, and for the no practice condition it was 9.83. These scores reflect inaccuracy two and three times more than that of the high achievers.

As Hacker, Bol, & Keener (2010) reported there have been many studies that have established a relationship between achievement level and calibration accuracy. Generally, high achievers are more accurate than low achievers yet often underconfident, and low achievers are usually less accurate than high achievers, yet often overconfident in their accuracy predictions (Hacker & Bol, 2004; Pajares & Graham, 1999). A notable exception was research with middle school math students (Bol et al., 2010). Both higher and lower achieving students were overconfident, although the higher achieving students were less so.

Despite some exceptions, the tendency for low achieving students to be overconfident and inaccurate has been replicated in the literature (Bol & Hacker 2000; Grimes 2002; Hacker et al., 2000). Kruger and Dunning term this the *unskilled but unaware* effect (1999). The results of the current research add to the already considerable body of evidence. A significant difference was found in calibration accuracy between high and low achievers. High achievers were more accurate overall. There was no difference between the high achievers in the practice group and no practice group in terms of accuracy, however there was a significant interaction; low achievers in the no practice condition were more accurate than low achievers in the practice condition group by 5 points on average. In other words, it seems that practice calibrating their test scores reduced accuracy among lower achievers rather than improving it. One possible explanation is that students in the practice condition believed they would do better on the SOL because of the practice they had calibrating their scores. This may have resulted in even more of a misalignment between judgment and performance on the final SOL Test.

### Self-Efficacy and Performance

Prior studies have shown that self-efficacy is a predictor of achievement, perhaps because students with higher self-efficacy, regardless of achievement level, tend to work longer at solving problems (Schunk, 1991; Zimmerman, 2006). However, in the current study, it was hypothesized that self-efficacy would not necessarily predict performance but would predict bias. Evidence indicates that lower achieving students often express overconfidence incongruent with their actual performance (Kruger & Dunning, 1999). An achievement paradox occurs when individuals possess high self-efficacy but are low achievers. This was true of many of the students in this sample. Part of the hypothesis proved to be true; self-efficacy did not predict SOL performance. Students had self-efficacy scores slightly above the middle of the scale. The mean student self-efficacy score was 3.5 on the five point scale. A score of 3.5, although not high, represents more positive than negative self-efficacy. Overall, student performance did not match their self-efficacy beliefs considering that the average score on the SOL for the students who took the self-efficacy scale was 31.4 out of 50, equating to 63%.

This is dissimilar to Chen's (2003) findings, where a strong relationship was found in a correlational analysis between self-efficacy and math performance. Chen's regression analysis revealed that self-efficacy predicted 25.4% of the variance in math performance. Perhaps self-efficacy is a better predictor of performance when the students are higher achieving. Chen's 7<sup>th</sup> grade students on average ranged between the 65% and 70% percentile marks on the Iowa Test of Basic Skills (ITBS) Math, a norm-referenced assessment. The idea that higher achieving students' self-efficacy scores are better

predictors of performance seems logical considering that higher achieving students also tend to be better calibrated than lower achieving students. This may translate to better self-evaluation skills. However, this argument is speculation and requires further study.

Brookhart et al (2006) examined the dynamics of effort and motivation among 8<sup>th</sup> grade science and social studies students and reported findings similar to Chen's. The researchers analyzed the effects of the classroom assessment environment, self-efficacy, and effort. Of the three constructs, self-efficacy was the strongest predictor of achievement.

In a study more similar to the current research, Pajares and Miller's (1997) 8<sup>th</sup> grade algebra students measured self-efficacy at global and local levels. Recall that global measures ask students to predict overall scores on a task while local measurements ask students to predict outcomes for individual items. Students were given the tests and asked to rate their self-efficacy about solving math problems after looking at the problems but not solving them (global) and then again for each item on two tests (local); one was a multiple choice format and the other a performance, open-ended format. The researchers found that student self-efficacy judgments did not differ according to test format even though students performed worse on the open-ended test formats. This led the authors to conclude that students may be even less well calibrated and more overconfident about their math abilities than expected. They may guess some answers correctly on multiple choice tests, thereby boosting their score, and making their calibration judgments appear less inflated. On the open-ended tests, guessing is not possible. Students performed worse on these items and their overconfidence was



subsequently more pronounced. The authors posited the idea that students may expect multiple choice assessments and therefore base their self-efficacy judgments on their performance on multiple choice tests regardless of test format. Therefore, the authors suggested that the predictive utility of self-efficacy is altered depending on the assessment format. However, unlike the present study, self-efficacy did predict performance.

The difference in findings between the present study and other studies linking self-efficacy and achievement warrants further investigation. While a relationship was not found between self-efficacy and performance, it could be that the sample was just too small in this study and lacked statistical power. It may be fruitful to replicate this research with more students of varying levels of achievement to determine if self-efficacy is a stronger predictor of achievement for higher achieving students compared to lower achieving students.

#### Self-Efficacy and Calibration Accuracy

The research that is available indicates there is a relationship between self-efficacy and calibration accuracy (Bembenutty, 2009). Some of the existing literature suggests a positive relationship. That is, students who are better calibrated have high self-efficacy (Bembenutty, 2009; Pajares & Miller, 1997). Chen's research with 7<sup>th</sup> grade students supported this contention. Students who were better calibrated had higher levels of self-efficacy; however, it was dependent on the difficulty level of the problem (Chen, 2003). The hypothesis for the present study was that self-efficacy would predict calibration bias. It was anticipated that an inverse relationship would exist between self-

efficacy and calibration accuracy, that higher self-efficacy scores would predict calibration inaccuracy and overconfidence. This expectation was derived in part from the repeated findings of other studies (Bol & Hacker, 2001; Grimes, 2002; Hacker et al., 2000) showing that higher achieving students were often underconfident in their predictions, and lower achieving students were often overconfident. Because the present research was conducted with lower achieving students rather than higher achieving students it was anticipated that students might have overly positive self-efficacy beliefs and poor calibration skills.

Reinforcing the argument that self-efficacy might be inversely related to calibration accuracy, Pajares and Miller's study (1997) found no difference in perceived self-efficacy between the prealgebra and algebra groups even though the algebra group outperformed the prealgebra on the performance measure. Both groups had a mean confidence rating of 80%; they believed they could solve 80% of the problems. This indicates that the prealgebra students were more overconfident in their predictions than the algebra group. They had the same expectations but performed worse.

In 2002, Klassen reviewed 22 articles exploring the self-efficacy beliefs of learning disabled (LD) students. In five of six studies investigating writing skills, LD students expressed overconfidence, even though students had been identified as having writing disabilities. However, in the five studies investigating math skills and self-efficacy beliefs, only one study found overestimates of efficacy beliefs (p. 97).

The research regarding the relationship between self-efficacy and calibration accuracy is limited, and somewhat contradictory. This study did not help clarify the

question because our hypothesis that self-efficacy would predict bias was not found. No relationship was found between self-efficacy and calibration accuracy. Perhaps one reason why self-efficacy did not predict accuracy or bias was that students in this study responded to the self-efficacy scale at the beginning of the semester, before they were familiar with the content of the tests. In addition, the sample size may have been too small to detect effects. Also, the self-efficacy scale was global rather than local. That is, students answered prompts about how well they would do in math as a subject rather than on individual math topics. Perhaps if the self-efficacy scale was introduced at a local level, based on different topics, there would be more of a relationship.

As discussed earlier, self-efficacy is a contextual domain. Individuals have different self-efficacy beliefs dependent on the task (e.g. math, English, or riding a bike). It would be informative to replicate this study using a topical self-efficacy scale that asked students to rate their efficacy beliefs dependent on the type of problem (e.g. geometry or statistics), as well as on a global scale. While this is not feasible for the high stakes test event itself, it would be possible to have students to provide a self-efficacy rating for each topic in addition to their calibration predictions.

#### Student Perceptions of Calibration

It was hypothesized that students would find topical calibration practice helpful in increasing their accuracy and performance. This was marginally supported. Recall that students were asked to answer three questions about the efficacy of the calibration strategy towards increasing their performance. The first question asked them to consider whether calibration practice helped them think about how to do better on the next test; the

second question asked them if calibration practice helped them think about what areas they needed to work on to do better; and the third question asked them if calibration practice helped them do better on their tests. Students were more positive about questions one and two. Seventy-five percent of the students either agreed or strongly agreed with the first question, and eighty-three percent agreed or strongly agreed with the second question. In contrast, only fifty-four percent agreed or strongly agreed to the third question. The first two questions were about metacognition, whether practice predicting their scores helped the students think about how to do better on the next test and what areas they needed to work on specifically. The third question was a direct question asking the students if the practice with calibration helped them to do better on the tests. Their perceptions seem to indicate that although the calibration strategy did make them think more about their test scores it did not translate to better scores for them.

Other studies have been more successful in terms of student beliefs regarding calibration strategies (Brookhart et al., 2004). Third grade students were asked to make predictions on reflection sheets of how well they would do on weekly timed multiplication tests and graph their predictions and results. Although students were not asked directly about their perceptions, researcher interviews with participating teachers at the end of the study indicated that students enjoyed participating in the self-assessment and seeing their progress on the graphs as their accuracy increased (p. 225).

Although anecdotal, Cleary reported (2009) positive student perceptions using the Self Regulation Empowerment Program (SREP). The SREP uses strategies designed to reduce the gap between self-judgments and actual performance. Cleary specifically

recalled a student who was initially resistant to participating in the SREP but became more open to learning about calibration accuracy strategies after seeing the 27-point difference between her predicted and actual test scores (p.168).

Hacker et al. (2009) suggest that future research into calibration include qualitative data, as most of the existing literature is quantitative. The literature that is available regarding student perceptions of calibration strategies is especially scant. It would be fruitful to investigate student perceptions in future studies to determine what students believe to be the most efficacious strategies in helping them develop metacognitive skills that will translate to improved performance and accuracy. Although students' positive perceptions of the efficacy of the topical calibration strategy are encouraging, the present results suggested that students thought it helped when it did not, so caution is needed.

### Limitations

There were several constraints that limited the validity of this study. Selection bias was a potential threat to internal validity if the groups differed on important variables before the treatment was implemented. It was not possible to determine if significant differences existed at the onset, although all participants were in the same grade level and course. It was also possible that there were teacher effects arising from having different teachers for the two conditions. However, students were taught the same material, at the same time, and in the same way, to ensure that students were prepared for tests. Treatment fidelity and diffusion may have compromised internal validity if students discussed the study between groups outside of class, especially with the group that comprised the no practice condition. Attrition also threatened internal

validity because it was difficult to have enough cases to match and track due to absences for different tests. Other threats to validity related to the self-report measures. Students may not have taken the study seriously, or may not have wanted to take the time necessary to reflect accurately on their predictions, possibly resulting in higher inaccuracy.

Sampling bias affected both internal and external validity, as the sampling for this study was limited by finding schools and teachers willing to participate. Students were all from one urban, middle school and only one subject and grade were used. Although this school district was purposefully selected because it was an urban school with lower achieving students, that also limits external validity. Sample size was also a threat to external validity. Only 54 students participated in the practice condition. Some data had to be omitted because there were too many cases where students were absent for one of more of the days the topical predictions were made. In addition, although the study took place over the course of three months, only three testing occasions were used for practice. More practice may have yielded different results.

As mentioned earlier, the topical strategy practice may not have been a robust enough to result in substantial gains in accuracy or performance. The addition of prompts requiring more direct reflection, such as calibration guidelines may have increased the efficacy of the study.

#### Future Directions for Research

Little research is available regarding the effectiveness of metacognitive strategies as calibration practice for urban, public school students. Although the current study was

not effective in increasing accuracy or achievement, it is likely that the results are inconclusive because of the limitations of the study. More studies that introduce calibration strategies to this group of students, especially in math, are needed. Research does indicate that this group has a misalignment between accuracy and performance (Bol et al., 2010).

Cleary (2009) suggested that it was highly likely that enabling students to strategically reflect on their calibration about test performance would have a positive impact, but that more research was necessary. This study attempted to address this by asking students to predict their test scores for each topic on their tests, with the expectation that this additional reflection would encourage more understanding of specific strengths and weaknesses, thereby increasing performance. However, this was not the case. As mentioned earlier, perhaps the calibration strategy was not powerful enough to elicit much reflection, and more specific reflection prompts need to be included. It may be fruitful to include calibration guidelines (Bol, et al., 2010), or graphing strategies (Zimmerman, 2006) in future research.

Another direction would be to determine if self-efficacy is more predictive when achievement level is considered. Other studies have found a positive relationship between self-efficacy and achievement; high self-efficacy predicts high achievement (Zimmerman et al., 2006), however it may be beneficial to see if a relationship exists between self-efficacy and performance when achievement level is varied.

Self-efficacy was not found to be predictive of achievement or accuracy in this study when it was measured globally at the beginning of the data collection. However,

asking students to rate their self-efficacy on each topic may increase its predictive power, and would be a logical next step in this research. The self-efficacy scale for this study was based on math as a global subject, rather than individual math subjects. Since self-efficacy is domain-specific (Zimmerman et al., 2006), perhaps increasing the specificity of the scale to math topics rather than math as a whole subject would increase the validity of the scale, and become a more effective research tool for inquiry and intervention.

### Educational Implications

In much of the literature it has been suggested that teachers incorporate metacognitive strategies such as calibration practice into their classroom assessment practices, especially for younger students and those at lower levels of academic achievement (Cleary, 2009; Nietfeld et al., 2006; Pajares et al., 1997) Though this study did not support the efficacy of a topical calibration strategy, others have had more success. Bol's study (2010) with calibration guidelines and Brookhart et al.'s (2004) success with graphing calibration predictions and results are evidence that calibration strategies do work. More studies are needed to determine if these, or others are the best strategies for improving accuracy and performance.

Another consideration is what scaffolding is needed once a strategy has been vetted and students are made aware of the discrepancy between their calibration judgments and actual performance. It is necessary that students understand how to apply the knowledge in ways that will help them improve their performance on tests. Teachers may need to instruct them to direct their focus appropriately (e.g. on specific topics where there was a misalignment between their expectations and outcomes). Low-achieving



students in urban settings may especially need this focus to help them alter their test-taking performance so that they better prepare for tests armed with an understanding of their strengths and weaknesses. This study and previous research makes it clear that higher performing students are more proficient at predicting their performance on these tests (Bol et al, 2010; Chen, 2003; Walck, 2010). The students in need of intervention targeting metacognition and calibration strategies are lower-achieving students, especially those faced with high stakes testing situations. Hacker et al.'s study (2008) demonstrated that extrinsic rewards were more effective at increasing calibration postdiction accuracy amongst lowest achieving students than a reflection intervention. Helping these students better align their calibration judgments may help them direct their attention towards those areas they are weakest in, thereby increasing their performance.

### Summary and Conclusions

This study focused on the use of a topical calibration strategy to increase calibration accuracy and improve test performance. Students were regular sixth grade math students at an urban middle school. Students that participated in the practice condition had the opportunity to practice calibrating their scores topically for three tests and then the final high-stakes SOL Test.

The research questions that were investigated included the effect of the topical calibration strategy on test performance and calibration accuracy, the impact of achievement level on calibration accuracy, and the use of self-efficacy as a predictor for achievement or accuracy, and student perceptions of the topical calibration practice.

The topical calibration strategy did not increase test performance or calibration accuracy in the practice condition. Performance and accuracy scores for students in the practice condition were similar to students in the no practice condition. As found with prior research (e.g. Bol et al, 2010; Hacker et al., 2000; Hacker et al., 2010; Kruger et al, 1999), accuracy was significantly impacted by achievement level; high achievers were more accurate in their calibration accuracy than low achievers.

Self-efficacy was not found to be a predictor of either achievement or calibration accuracy. However, this contradicts other research indicating that there is a link between self-efficacy and achievement (Chen, 2003; Pajares et al., 1997). Further research is needed to help clarify this relationship.

Student perceptions of the effectiveness of this strategy were between neutral and slightly positive. Their answers indicated that the strategy was more valuable as a metacognitive scaffold; it helped them consider areas they needed to work on rather than as a tool that increased their test performance.

Although the current study did not improve performance on high stakes tests or improve calibration accuracy as was hoped, it did provide further insight into the metacognitive processes of urban, public school students and supports the need for further research with this population. These students are in need of calibration strategies that can be used to minimize the misalignment that exists in many of these students' beliefs about their performance and their actual performance.

## REFERENCES

- Bandura, A. (1989) Human agency is social cognitive theory. *American Psychologist*, 44, 1175-1184.
- Bandura, A. (2006). Adolescent development from an agentic perspective. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.339-67). Greenwich, CT: Information Age Publishing.
- Barksdale-Ladd, M. A. & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education*, 51, 384-97.
- Bembenutty, H. (2009). Three essential components of college teaching: Achievement calibration, self-efficacy, and self-regulation. *College Student Journal*, 43(2), 562-570.
- Boekaerts, M. & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction*, 20, 372-382.
- Bol, L. (2004). Teachers' assessment practices in a high-stakes testing environment. *Teacher Education and Practice*, 17, 162-181.
- Bol, L. & Garner, J. K. (2011) Challenges in supporting self-regulation in distance education environments. *Journal of Computing in Higher Education*, 23, 104-123.
- Bol, L. & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education*, 69, 133-51.

- Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, 73, 269-90.
- Bol, L. & Nunnery, J.A. (2004). The impact of high-stakes testing on restructuring efforts in schools serving at risk students (pp. 101-117). In G. Taylor (Ed.), *In pursuit of equity and excellence: The educational testing and assessment of diverse learners*. Lewiston, New York: Edwin Mellon Press.
- Bol, L., Riggs, R., Hacker, D. J., Dickerson, D. & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education*, 21, 81-96.
- Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology*.
- Bong, M. (2006). Asking the right question: How confident are you that you could successfully perform these tasks? In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.339-67). Greenwich, CT: Information Age Publishing.
- Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics*, 57, 213-227.
- Brookhart, S.M., Walsh, J.M., & Zientarski, W. A. (2006). The dynamics of motivation and effort for classroom assessments in middle school science and social studies. *Applied Measurement in Education*, 19(2), 151-184.

- Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 3, 245-281.
- Center for Education Policy (2009). *Are achievement gaps closing and is achievement rising for all? State test score trends through 2007-08, part 3*. Washington, D.C.
- Chen, P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences*, 14, 79-92.
- Chen, P. & Zimmerman, B. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *The Journal of Experiential Education*, 75, 221-44.
- Cleary, T. J. (2009). Monitoring trends and accuracy of self-efficacy beliefs during interventions: Advantages and potential applications to school-based settings. *Psychology in the Schools*, 46(2), 154-171.
- Dembo, M. H. & Eaton, M. J. (2000). Self-regulation of academic learning in middle-level schools. *The Elementary School Journal*, 100, 473-90.
- Flannelly, L. T. (2001). Using feedback to reduce students' judgment bias on test questions. *Journal of Nursing Education*, 40, 10-16.
- Grabe, M. & Flannery, K. (2010). A preliminary exploration of on-line study question performance and response certitude as predictors of future examination performance. *Journal of Educational Technology Systems*, 38(4), 457-472.
- Grimes, P. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *The Journal of Economic Education*, 33(1), 15-30.

- Haberman, M. (2004). *Urban education: The state of urban schooling at the start of the 21<sup>st</sup> Century*. Haberman Educational Foundation, Houston, TX.
- Hacker, D. J., Bol, L., Horgan, D. & Rakow, E. A. (2000) Test prediction and performance in a classroom context *Journal of Educational Psychology*, 92, 160-70.
- Hacker, D. J. & Bol, L. (2004). Metacognitive theory considering the social cognitive influences. In D. M. McInerney & S. Van Etten (Eds.), *Big Theories Revisited* (p.275-97). Greenwich, CT: Information Age Publishing.
- Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, 3, 101-121.
- Hacker, D. J., & Bol, L. (2011, April). *Comparing absolute and relative accuracy in a classroom context*. Paper presented at Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Hacker, D.J., Bol, L., & Keener, M. C. (2009). Metacognition in education: A focus on calibration. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metgacognition in Education* (p. 429-455). New York, NY: Routledge.
- Hoffman, L. M. & Nottis, K. E. K. (2008). Middle school students' perceptions of effective motivation and preparation for high-stakes tests. *NASSP Bulletin*, 92, 209-23. Retrieved February 19, 2010 from <http://online.sagepub.com>.
- Horgan, D. (1990). Student's predictions of test grades: Calibration and metacognition. In *American Educational Research Association Annual Meeting*. Boston, MA.

- Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, 4, 161-176.
- Klassen, R. M. (2002). A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly*, 25(2), 188-202.
- Klassen, R. M. (2006). Too much confidence? The self-efficacy of adolescents with learning disabilities. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.339-67), Greenwich, CT: Information Age Publishing.
- Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the performance of business students on examinations. *Journal of Education for Business*, 79(4), 217-24.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it; how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121-1134.
- Lieberman, V. (2004). Local and global judgments of confidence. *Journal of Experimental Psychology*, 30(3), 729-732.
- Lin-Miao, L., & Zabucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345-391.
- Linn, R. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, 7, 29-38.
- Lynn, J. (2008). STEM: The 21<sup>st</sup> Century Sputnik. *Kappa Delta Pi*, 4, 152-153.

- Maki, R. H., Shields, M., Wheeler, A. E. & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology*, 97, 723-31.
- Mathis, W. J. (2003). *No child left behind: What are the costs? Will we realize any benefits?* (Report No. ED477646). Evaluative Report. Retrieved from ERIC database (EA032551).
- Midgley, C, Kaplan A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L.H., Anderman, E., & Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*. 23, 113-131.
- Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J., Nelson, J., Roeser, R. & Urdan, T. (2000). Manual for the Patterns of Adaptive Learning Scales. University of Michigan. Retrieved February 3, 2011 from: <http://www.umich.edu/~pals/>
- Miller, D. C., Anindita, S., Malley, L. B., and Burns, S.B. (2009). Comparative indicators of education in the United States and other G-8 countries: NCES 2009-039. National Center for Education Statistics.
- Morrison, J., & Bartlett, R. V. B. (2009). STEM as a curriculum an experiential approach. *Education Week*, 2.



- 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*(1), 30-38.
- Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7-28.
- Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159-79.
- Pajares, F. (2006). Self-efficacy during childhood and adolescence: Implications for teachers and parents. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.339-67), Greenwich, CT: Information Age Publishing.
- Pajares, F. & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24*, 124-39.
- 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*, 213-28.
- Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for struggling low-income urban students: Evidence from Massachusetts. *Educational and Evaluation Policy Analysis, 32*(1), 5-23.

- Popham, W.J. (2001). Uses and misuses of standardized tests. *NASSP Bulletin* 85(6), 24-31.
- Puncochar, J., & Fox, P. W. (2004). Confidence in individual and group decision-making: When "two heads" are worse than one. *Journal of Educational Psychology*, 96, 582-591.
- Ramdass, D. H. (2009). Improving fifth grade students' mathematics self-efficacy calibration and performance through self-regulation training. (Doctoral dissertation, City University of New York, New York, 2009). *Dissertations & Theses: Full Text*, 70(07).
- Roebbers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *The British Journal of Educational Psychology*, 79, 749-67.
- Ryan, K. E., Ryan, A. M. & Arbuthnot, K. (2007). Students' motivation for standardized math exams. *Educational Researcher*, 36, 5-13.
- Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology*, 19(2), 143-154.
- Schraw, G. (2009). Measuring metacognitive judgments. (pp. 415-429). In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.) *Handbook of Metacognition in Education*. New York, NY: Routledge.
- Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, 26 (3 & 4), 207-31.

- Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review*, 20, 463-467.
- Schunk, D. H. & Meece, J. L. (2006). Self-efficacy development in adolescence. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.71-96), Greenwich, CT: Information Age Publishing.
- Spock, B. & Parker, S. (1998). *Dr. Spock's Baby and Childcare* (7<sup>th</sup> ed.). New York, NY: Pocket Books.
- Stolp, S. & Zabucky, K. M. (2009). Contributions of metacognitive and self-regulated learning theories to investigations of calibration of comprehension. *International Electronic Journal of Elementary Education*, 2(1), 7-31.
- Swanson, C. B. (2004). The new math on graduation rates. *Education Week*, 23, p.30-31.
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95, 66-73.
- U.S. Department of Education. (2011). National Center for Education Statistics <http://nces.ed.gov/>.
- US Department of Education Institute of Education Sciences. (2009). Comparative Indicators of Education in the United States and Other G-8 Countries. Retrieved February 4, 2010, <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009039>

Virginia Department of Education, Commonwealth of Virginia. (2010). *Accountability*

*Guide*. Retrieved January 29, 2010, from

[http://www.doe.virginia.gov/statistics\\_reports/school\\_report\\_card/accountability\\_guide.shtml](http://www.doe.virginia.gov/statistics_reports/school_report_card/accountability_guide.shtml)

Virginia Department of Education (2011). Accreditation and AYP Reports, Retrieved

January 5, 2012, from

[http://www.doe.virginia.gov/statistics\\_reports/accreditation\\_ayp\\_reports/title1/index.ex.shtml](http://www.doe.virginia.gov/statistics_reports/accreditation_ayp_reports/title1/index.ex.shtml)

Virginia Department of Education (2011). *Enrollment and Demographics*. Retrieved

January 29, 2010, from

[http://www.doe.virginia.gov/statistics\\_reports/enrollment/fall\\_membership/index.shtml](http://www.doe.virginia.gov/statistics_reports/enrollment/fall_membership/index.shtml)

Walker, C. O. & Greene, B. A. (2009). The relations between student motivational beliefs and cognitive engagement in high school. *The Journal of Educational Research*, 102, 463-71.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology*, 27, 551-572.

Zimmerman, B. J. & Cleary, T. J. (2006). Adolescents' development of personal agency: The role of self-efficacy beliefs and self-regulatory skill. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (p.339-67), Greenwich, CT: Information Age Publishing.

**APPENDIX A: TEST CALIBRATION BY TOPIC (PRACTICE CONDITION ONLY)**

Student Name \_\_\_\_\_ Teacher \_\_\_\_\_

**BEFORE THE TEST:** Look at each of the topics listed in the table below, and the number of problems there will be in each category. Guess how many questions you will answer correctly for each topic and write that number in the appropriate box in the “BEFORE the test” column.

**AFTER THE TEST:** When you get your scored test back, list how many problems you answered correctly for each topic in the “AFTER the test” column.

Test Calibration by Topic

Qtrly Test Number of SCORED questions for each topic	BEFORE QUARTERLY TEST	AFTER QUARTERLY TEST	Unit Test Number of SCORED questions for each topic	BEFORE UNIT CFA TEST	AFTER UNIT TEST	MOCK SOL Number of SCORED questions for each topic	BEFORE MOCK SOL TEST	AFTER MOCK SOL TEST
	PREDICTED SCORE	ACTUAL SCORE		PREDICTED SCORE	ACTUAL SCORE		PREDICTED SCORE	ACTUAL SCORE
Numbers & Number Sense 8			Identify Integers 8			Numbers & Number Sense 8		
Computation & Estimation 10			Compare Integers 8			Computation & Estimation 10		
Measurement & Geometry 12			Solving Equations 4			Measurement & Geometry 12		
Probability & Statistics 8						Probability & Statistics 8		
Patterns & Functions 12						Patterns & Functions 12		
Total Number of Items  50			20			50		

## APPENDIX B: STUDENT OPT-OUT FORM

### Parent Opt-Out Permission Form Math Self Efficacy and Prediction Study

Certain 6<sup>th</sup>, 7<sup>th</sup>, and 8<sup>th</sup> grade math classes at Northside Middle School are participating in a Math Self Efficacy and Prediction Study by a P.H.D candidate researcher at Old Dominion University. Students will be asked to predict their grades on tests, including the SOL exams, and then record their actual grades next to their predictions. Students that participate will also be asked to answer questions about how they perceive their abilities in math class. The purpose of the study is to heighten student self-awareness of their performance on math tests, and increase their own self-efficacy beliefs.

Students will put their names, race, and gender on the survey and prediction worksheets in order to track their progress to see if it improves, but identifying information will be removed and destroyed at the end of the study. This study will cause no risk to your child. Your child may benefit from the study through increased self-efficacy beliefs and improvement on their test scores.

We would like all students in the selected math classes to take part in the study, however, participation is voluntary—no action will be taken against the school, you, or your child, if your child does not participate.

**This form is only returned if you do NOT want your child to participate.**

**If you do NOT want your child to participate in the study:**

**Please check the box below, sign, and have your child return the form to their math teacher no later than March 4, 2011.**

If you have any questions about the study, please call or email: Rose Riggs, 757-319-8692, rrigg003@odu.edu.

**[    ] My child may NOT participate in the Self Efficacy and Prediction Study.**

Student's name: \_\_\_\_\_ Grade: \_\_\_\_\_

Parent's Name: \_\_\_\_\_

Parent's signature: \_\_\_\_\_ Date: \_\_\_\_\_

## APPENDIX C: PATTERNS OF ADAPTIVE LEARNING SCALES (PALS)

## Academic Self Efficacy

Student Name \_\_\_\_\_ Teacher \_\_\_\_\_

**Directions:** Please answer each of the questions below using the scale provided.

1. I'm certain I can master the math skills taught in class this semester

1                      2                      3                      4                      5

NOT AT ALL TRUE              SOMEWHAT TRUE              VERY TRUE

2. I'm certain I can figure out how to do the most difficult math class work.

1                      2                      3                      4                      5

NOT AT ALL TRUE              SOMEWHAT TRUE              VERY TRUE

3. I can do almost all the work in math class if I don't give up.

1                      2                      3                      4                      5

NOT AT ALL TRUE              SOMEWHAT TRUE              VERY TRUE

4. Even if the work in math class is hard, I can learn it.

1                      2                      3                      4                      5

NOT AT ALL TRUE              SOMEWHAT TRUE              VERY TRUE

5. I can do even the hardest work in math class if I try.

1                      2                      3                      4                      5

NOT AT ALL TRUE              SOMEWHAT TRUE              VERY TRUE

## APPENDIX D: INSTRUCTIONS FOR SELF-EFFICACY SCALE PROCTORING

**BEFORE STUDY INSTRUCTIONS:** Write the question below, as well as the answer scale, on the board. Then read the following script to the students:

I am going to give you a survey in a minute, which is a series of questions, to find out how you feel about your ability to do your work in this math. The survey is not a test and there are no right or wrong answers; the information that is collected will remain confidential and your parents and peers will not see your specific responses to any of the questions. Some questions may sound very similar to others in the survey, but this is important to make sure we have a good idea of what you think. I will review a sample question first to familiarize you with the format of the questions.

An example of a question that is like the ones on the survey is written on the board:

**“It’s important to me that I learn a lot of new concepts in math this year.”**

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>NOT AT ALL TRUE</b>		<b>SOMEWHAT TRUE</b>		<b>VERY TRUE</b>

If you thought this statement was not at all true of you, you would choose 1, if you thought it was somewhat true of you, you would choose 3, and if you thought it was very true of you, you would choose 5. You can also choose a number in between, either 2 or 4, if you feel that your answer is somewhere in between these categories.

Are there any questions?



## APPENDIX E: TEACHER INSTRUCTIONS TO STUDENTS

### **Topical Test Prediction Practice Participants:**

Please read this to your students prior to the first practice test.

*"Based on research we think it helps students do better on tests when they guess, or predict, their grade in advance of the test. An Old Dominion University research student has asked you to help her see if doing this might improve your scores. You will predict your grades for unit and quarterly tests, as well as the Standards of Learning/Standards of Learning End-of-Course exam. I will be passing out a form prior to each of these tests for you to write your predictions on, and then returning the form to you for you to write your actual score when you get the test back. Please do your best and be honest in your predictions. This won't count toward your grades. Your answers and identity will be kept secret."*

(Provide an example)

*"For example, the first test you will take has 50 problems, divided into 4 or 5 different topics, such as Statistics. In the "Before the Test "column, in the Statistics row, enter the number of problems you think you will answer correctly for that topic. "*

*"Then after taking the test, you will write how many questions you did answer correctly for that topic in the next column. "*

*"Are there any questions?"*

### **SOL ONLY Topical Test Prediction Students ONLY:**

*"Based on research we think it helps students do better on tests when they guess, or predict, their grade in advance of the test. An Old Dominion University research student has asked you to help her see if doing this might improve your scores. Please do your best and be honest in your predictions. This won't count toward your grades. Your answers and identity will be kept secret. You are being asked to predict how many problems you expect to answer correctly for each section on the Standards of Learning Test. The test has 50 problems, divided into 4 or 5 different topics, such as Statistics. In the "Before the Test "column, in each row, enter the number of problems you think you will answer correctly for that topic. For example, if there were 8 Statistics problems, and you think you will get all 8 right, then enter 8 in the 'Statistics' row."*

# APPENDIX F: FINAL SOL TEST CALIBRATION BY TOPIC (ALL STUDENTS)

Student Name \_\_\_\_\_ Teacher Name \_\_\_\_\_

**BEFORE THE TEST:** Look at each of the topics listed in the table below, and the number of problems there will be in each category. Guess how many questions you will answer correctly for each topic and write that number in the appropriate box in the “BEFORE the test” column.

**AFTER THE TEST:** When you get your scored test back, list how many problems you answered correctly for each topic in the “AFTER the test” column.

EOC SOL Test Topics	Total Number of SCORED Problems for each topic	BEFORE the test	AFTER the test
		PREDICTED SCORE (How many problems I expect to answer correctly)	ACTUAL SCORE (How many problems I did answer correctly)
Number & Number Sense	8		
Computation & Estimation	10		
Measurement & Geometry	12		
Probability and Statistics	8		
Patterns, Functions, & Algebra	12		
Total Items	50		

## APPENDIX G: STUDENT PERCEPTIONS OF STRATEGY QUESTIONS

Student Name \_\_\_\_\_ Teacher \_\_\_\_\_

1. Practice in predicting my overall test scores helped me think about how to do better on the next test. (Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

2. Practice in predicting how well I would do helped me think about what areas I needed to work on to do better on the next test. (Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

3. Practice predicting my test scores helped me to do better on my tests.  
(Circle your answer.)

Strongly Disagree      Disagree      Agree      Strongly Agree

## VITA

Rose Riggs

155 Lembla Street  
Norfolk, Virginia 23503  
Email: [rrigg003@odu.edu](mailto:rrigg003@odu.edu)  
Cell Phone: 757-319-8692 Home Phone: 757-275-7346

### EDUCATION:

- August 2012 Expected date of completion  
Ph.D. in Educational Curriculum and Instruction, Old Dominion University, Norfolk, Virginia  
Dissertation: Can Practice Calibrating By Test Topic Improve Public School Students' Calibration Accuracy And Performance On Tests?
- 2004 M.S. in Secondary Education, English, Old Dominion University, Norfolk Virginia  
Concentration: Teacher Licensure
- 1988 B.S. in Business Administration, Utica College of Syracuse University, Utica, New York  
Concentration: Personnel Management

### EXPERIENCE:

- 2008-present Senior Instructional Designer, Naval Education and Training Command, Norfolk, Virginia
- 2006-2008 Senior Instructional Designer, AMSEC, Newport News, Virginia
- 2004-2006 Graduate Teaching Assistant, Old Dominion University, Norfolk, Virginia
- 2003-2004 English Teacher, Norfolk Public Schools, Norfolk, Virginia

### PUBLICATIONS/PRESENTATIONS

- Bol, L., Riggs, R., Hacker, D.J., Dickerson, D., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education*, 21, 81-96.
- Riggs, R., Suich K., Geggis, L. (2012). What's in your seabag? In The Center for Combat Stress Operational Stress Conference. San Diego, CA.