

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Calibration of Confidence Judgments in Elementary Mathematics: Measurement, Development, and Improvement

Permalink

<https://escholarship.org/uc/item/99z17038>

Author

Rutherford, Teomara

Publication Date

2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Calibration of Confidence Judgments in Elementary Mathematics:
Measurement, Development, and Improvement

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Teomara Rutherford

Dissertation Committee:
Professor George Farkas, Chair
Distinguished Professor Greg J. Duncan
Founding Dean and Professor Deborah Lowe Vandell
Distinguished Professor Jacquelynne Eccles

2014

DEDICATION
To
Ainsley and Corinna

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGEMENTS	vii
CURRICULUM VITAE	viii
ABSTRACT OF THE DISSERTATION	xv
CHAPTER 1: Introduction and Literature Review	1
CHAPTER 2: Study 1: The Measurement of Calibration in Real Contexts	21
CHAPTER 3: Study 2: Within and Between Person Associations of Calibration and Achievement	64
CHAPTER 4: Study 3: Changes in Calibration: In Response to Intervention and as Related to Changes in Achievement	107
CHAPTER 5: Summary and Conclusions	166
APPENDIX A Supplementary Tables for Study 1	180
APPENDIX B Supplementary Figures and Tables for Study 2	191
APPENDIX C Supplementary Tables for Study 3	201

LIST OF FIGURES

		Page
Figure 1.1	Self-Regulated Learning Cycle	20
Figure 2.1	Illustration of Calculation of Item-by-Item as Compared to More Macro Levels of Calibration	49
Figure 2.2	2x2 Contingency Table Expression the Relations between Accuracy and Confidence	50
Figure 2.3	Quiz Questions Appearing in ST Math	51
Figure 2.4	Distribution of Combinations of Confidence and Accuracy within the Actual St Math Quiz Data	52
Figure 3.1	Model of Conscious Regulation	98
Figure 3.2	Interaction Slopes, Sensitivity	99
Figure 3.3	Interaction Slopes, Specificity	100
Figure 4.1	Screen Shot from an ST Math Quiz	148
Figure 4.2	ST Math Study Design	149

LIST OF TABLES

		Page
Table 2.1	Common Indices of Agreement from 2x2 Contingency Tables	53
Table 2.2	Comparison of Sample Descriptives to County and State	54
Table 2.3	Mean Number of Questions Answered & in Each of the Calibration Categories, Aggregated Quiz Questions	55
Table 2.4	Odds Ratios and Marginal Effects from Logistic Regression of Zeroes in Quadrants C and D by Total Number of Questions	56
Table 2.5	Demographic Information and Descriptive Statistics for Study Subsamples	57
Table 2.6	Ten Common Measures of Calibration Calculated for Entire Sample from All Available Data	58
Table 2.7	Means of Ten Measures of Calibration Eliminating (Left) and Accommodating (Right) Missing Data	59
Table 2.8	Correlations among Measures of Calibration, Pre and Posttest Accuracy and Pretest Confidence in Reduced Sample	60
Table 2.9	Regression of Posttest Accuracy (Percentage of Items Correct) on Pretest Calibration and Accuracy	61
Table 2.10	Replication Sample: Regression of Posttest Accuracy (Percentage of Items Correct) on Pretest Calibration and Accuracy for Ten Measures of Calibration	62
Table 3.1	Grade and Demographic Information of Study Students	101
Table 3.2	Quiz Accuracy and Calibration Measures, by Grade	102
Table 3.3	Correlations between Calibration and Accuracy Measures	103
Table 3.4	Student-Level Regressions of Posttest Accuracy on Pretest Accuracy and Calibration	104
Table 3.5	Results from Hierarchical Regressions of Post-test Accuracy on Pre-test Accuracy, Calibration, & Covariates	105
Table 3.6	Comparison of Effect Sizes for 2010 Data across Analysis Methods	106
Table 4.1	Sample Demographics, Divided by Treatment Group	150

Table 4.2a	Descriptive Statistics of Calibration and Achievement Variables, Third Grade by Treatment Group	151
Table 4.2b	Descriptive Statistics of Calibration and Achievement Variables, Fourth Grade by Treatment Group	152
Table 4.3	Correlations between Calibration Measures and Achievement Measures	153
Table 4.4	Effect of Early Treatment Group on Calibration for Place Value	154
Table 4.5	Effect of Early Treatment Group on Calibration for First Three Objectives Encountered in ST Math	155
Table 4.6	Effect of Early Treatment Group on Calibration Aggregated Across Entire Year	156
Table 4.7	Descriptive Statistics for Calibration and Accuracy Across Years, Early Treatment Group	157
Table 4.8	Standardized Regression Coefficients Compared Across Analyses and Samples: Association of Treatment Group and Measures of Calibration	158
Table 4.9	Sample Characteristic Comparison for Question 2	159
Table 4.10	Descriptive Statistics of Selected Starting and Ending Quiz Pairs	160
Table 4.11	Association between Calibration Gain and Posttest Performance Gain, Paired Quizzes	161
Table 4.12	Association between Calibration Gain and Math CST Gain	162
Table 4.13	Slopes for Improvement in Accuracy and Calibration over Time	163
Table 4.14	Association between Calibration Growth and Growth of Quiz Posttest Performance	164
Table 4.15	Association between Calibration Growth and End-of-Year Math CST Scores	165

ACKNOWLEDGEMENTS

I would like to thank the MIND Research Institute and Orange County Department of Education for the provision of the data and for being a valuable part of the ST Math IES study.

Additionally, thanks to Matthew Peterson, Matt Feldmann, Andrew Coulson and others at MIND for their willingness to collaborate, their innovations, and their belief in potential and ability to make a difference. Thanks to Stephanie Schneider and Lauren Duran for their help in data collection, communication with schools, and overall support to keep things running smoothly.

I am grateful to the many members of the UCI faculty who have offered academic guidance and support. The members of my committee: George Farkas, Greg Duncan, Deborah Vandell, and Jacquelynne Eccles, and my proposal committee: AnneMarie Conley and Elizabeth Loftus. Other faculty involved in the ST Math project: Elizabeth van Es and Margaret Burchinal. Susanne Jaeggi, Barbara Sarnecka, and Stephanie Reich, for collaboration and mentorship. Mark Warschauer, for his belief in me and for lively conversation. Michael Martinez, for setting me on my course and for continuing to be a voice of guidance, even in spirit.

My time at UCI was enriched by friendship with my classmates and colleagues, and I am thankful for their encouragement, their humor, and for letting me share their journey. C3, for being there every step of the way. My alphabet soup of lab groups: SEDL, CAMP, WMP, and DLL, for a place to share ideas and to feel included. CRCL (and inner CRCL), for being my research home. Jeneen, Jana, Arya, and the CRCL research assistants, for getting things done and done well. I would also like to thank friends outside of UCI for companionship and support, especially the families and actors of Musical Theatre Village for providing a creative outlet and a place where my family could do something fun together.

Thanks to my daughters for understanding that I sometimes had to work on weekends, and for providing hugs and cute notes when I did.

I would not have been able to undertake a PhD or conduct my research without the support of my husband, Scott. His belief in me and willingness to move our family across the country for five years with uncertain returns was a driving force in my determination to succeed.

Finally, I would like to acknowledge the funding support of the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090527 to the University of California, Irvine and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0808392. This dissertation content is solely the responsibility of the author and does not necessarily represent the views of the funding agencies.

CURRICULUM VITAE

Teomara Rutherford

EDUCATION

- Ph.D.** Learning, Cognition, and Development, University of California, Irvine,
School of Education, 2014
Dissertation: Calibration of Confidence Judgments in Elementary Mathematics:
Measurement, Development, and Improvement
Committee: George Farkas (Chair), Greg Duncan, Deborah Vandell, Jacquelynne Eccles
- M.A.** Learning, Cognition, and Development, University of California, Irvine,
School of Education, 2009-2012
Advisor: Michael E. Martinez
- J.D.** Boston University School of Law, *cum laude*, 2000-2003
G. Joseph Tauro Distinguished Scholar, *Boston University Law Review* article editor
- Matriculated, Non-Degree Program, University of Texas at Austin, 1998-1999
30 credits in upper-division psychology; neuroscience coursework and research
- B.S.** Elementary Education, Florida International University, 1995-1997
Concentration: Computers in the Classroom; Faculty Scholar Honors Program

PUBLISHED MANUSCRIPTS

- Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Graham, J., Kibrick, M.,...Martinez, M. E. (2014) A randomized trial of an elementary school mathematics software intervention: Spatial-Temporal (ST) Math. *Journal of Research on Educational Effectiveness*. doi: 10.1080/19345747.2013.856978
- Schenke, K., Rutherford, T., & Farkas, G. (2014) Alignment of game design features and state mathematics standards: Do results reflect intentions? *Computers & Education*. doi: 10.1016/j.compedu.2014.03.019
- Rutherford, T. (2013). Emotional well-being and discrepancies between child and parent educational expectations and aspirations in middle and high school. *International Journal of Adolescence and Youth* 1–17. doi:10.1080/02673843.2013.767742
- Tran, N. A., Schneider, S., Duran, L., Conley, A. M., Richland, L., Burchinal, M., Rutherford, T., Kibrick, M., Osborne, K., Coulson, A., Antenore, F., Daniels, A., & Martinez, M. E. (2012). The effects of mathematics instruction using spatial temporal cognition on teacher efficacy and instructional practices. *Computers in Human Behavior*, 28(2), 340-349.
- Rutherford, T., Lee, D. S., & Martinez, M. E. (2011). Gender, spatial ability, and high-stakes testing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3237-3242). Austin, TX: Cognitive Science Society.

MANUSCRIPTS SUBMITTED & UNDER REVIEW

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M., *Raising the stakes: How students' motivation for mathematics associates with high- and low-stakes test achievement*. Revise and Resubmit.

PEER-REVIEWED PRESENTATIONS

Zhamharyan, H.* & Rutherford, T. (2014, November). *Less effective executive functioning after being sleep deprived*. Paper accepted to the annual meeting of the Psychonomic Society, Long Beach, CA.

*indicates mentored undergraduate student

Simzar, R. M., Martinez, M., Sanabria, T., Rutherford, T., Domina, T., & Conley, A. M. (2014, August). *Student motivation for mathematics and high-stakes versus low-stakes test achievement*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Schenke, K., Rutherford, T. & Farkas, G. (2014, April). *Linking educational technology to standardized assessments: Game content and features*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Chang, A., Rutherford, T., & Farkas, G. (2014, April). *I can do it!: Expectancy as a mediator of the ST Math effect on math achievement*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Rutherford, T. & Farkas, G. (2014, April). *Evaluation of ST Math treatment effects for special populations and by length of implementation*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Long, J. J., Rutherford, T., van Es, B., & Farkas, G. (2014, April). *Understanding the relationship between ST Math teacher professional development and its impact on students*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Johnson, L. G.*, Rutherford, T., & Lee, D. S. (2013, May). *Association between working memory training games and assessments*. Paper presented at the 25th annual meeting of the Association for Psychological Science, Washington, D.C.

*indicates mentored undergraduate student

Rutherford, T., Lee, D. S., Schenke, K., Chang, A., Tran, C., Young, N. S., Conley, A. M.,...Martinez, M. E. (2013, April). *Brain Boost: Randomized trial of a program to enhance intelligence in elementary and middle school*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rutherford, T., Schenke, K., Conley, A. M., & Martinez, M. E. (2012, August). *The association between math test scores, math expectancy, and cognitive abilities*. Paper presented at the annual meeting of the American Psychological Association, Orlando, FL

Schenke, K., Chang, A., Rutherford, T., Lee, D. S., Tran, C., Hinga, B., & Martinez, M. E. (2012, August). *Development, implementation, and evaluation of Brain Boost: A model*

for modifying direct cognitive ability in an engaging after school environment. Paper presented at the annual meeting of the American Psychological Association, Orlando, FL

Forrester, L. D.*, Rutherford, T., & Martinez, M. E. (2012, March). *Using the right words for reasoning: Relationships between specific word use and inductive reasoning.* Paper presented at the biennial meeting of the Society for Research on Adolescence, Vancouver, British Columbia, Canada.
*indicates mentored undergraduate student

Rutherford, T., Burchinal, M., Farkas, G., Graham, J.D., Kibrick, M., Long, J.J.,...Martinez, M.E. (2012, April). *Main and differential effects of a computer-assisted math intervention.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

Sheppard, A, Safavian, N., Rutherford, T., Albarran, A. S., & Conley, A. M. (2012, April). *Social network analysis of communication patterns within professional learning communities.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

Rutherford, T. (2011, December). *Illustration of program evaluation: ST Math randomized field trial.* Paper presented at the annual meeting of the California Educational Research Association, Anaheim, CA.

Rutherford, T., Hinga, B., Chang, A., Conley, A. M., & Martinez, M. E. (2011, August). *The effect of ST Math software on standardized test scores via improvement in mathematics expectancy.* Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Rutherford, T., Lee, D. S., & Martinez, M. E. (2011, July). *Gender, spatial ability & high-stakes testing.* Paper presented at the annual meeting of the Cognitive Science Society, Boston, MA.

Long, J. J., Rutherford, T., Richland, L. E., Graham, J. D., Antenore, F., Coulson, A., & Martinez, M.E. (2011,May) *Fidelity of implementation in ST Math.* Paper presented at the Science of Student Success Conference of the Learning & the Brain Society, Chicago, IL.

Lee, D.S., Rutherford, T., Hinga, B., Graham, J. D., & Martinez, M. (2011,May) *Brain Boost: A model for direct cognitive enhancement.* Paper presented at the Science of Student Success Conference of the Learning & the Brain Society, Chicago, IL.

Rutherford, T., Graham, J. D., Kibrick, M., Burchinal, M., Lee, D. S., Long, J. J.,...Martinez, M. E. (2011, April). *Change in standardized test scores in response to an individualized math intervention.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Rutherford, T., Conley, A. M., & Karabenick, S. A. (2011, April). *Achievement goal orientations and preference for competitive careers.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Sheppard, A, Safavian, N., Rutherford, T., Albarran, A. S., & Conley, A. M. (2011, April). *A social network analysis of teachers' professional learning communities.* Paper presented

at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Rutherford, T., Conley, A. M., & Karabenick, S. A. (2011, March). *Motivationally blessed? Motivation and achievement among gifted and propensity score matched comparison group*. Paper presented at the annual meeting of the Society for Research in Child Development, Montreal, Canada.

Rutherford, T., Kibrick, M., Burchinal, M., Richland, L. E., Conley, A. M., Osborne, K.,...Martinez, M. E. (2010, May). *Spatial temporal mathematics at scale: An innovative and fully developed paradigm to boost math achievement among all learners*. Paper presented at the annual meeting of the American Educational Research Association, Denver CO.

Rutherford, T. (2010, June). *Emotional well-being and matches between child and parent educational aspirations and expectations*. Paper presented at the PSID CDS III New Directions Workshop, Ann Arbor, Michigan.

Kibrick, M., Rutherford, T., Burchinal, M., Richland, L. E., Conley, A. M., Long, J. J., ...Martinez, M. E. (2010, June). *The effects of ST Math on standardized test scores: A randomized field study*. Paper presented at the Fifth Annual IES Research Conference, Washington, DC.

PUBLICATIONS IN OTHER AREAS

Rutherford, T. (2009, October). Copyright the wrongs: Legal concepts every photographer should know. *Design Aglow, 1*(9). Available from <http://designaglow.com>

Rutherford, T., & Warneck, S. (2002) Offender-victim contact. In *Bench Book on Children and the Courts; Improving Court Responses to Child Victims of Intra-familial Violence and Sexual Abuse*. Boston, MA: Children's Law and Policy Initiative.

Rutherford, T., & Warneck, S. (2002) Non-offending guardians. In *Bench Book on Children and the Courts; Improving Court Responses to Child Victims of Intra-familial Violence and Sexual Abuse*. Boston, MA: Children's Law and Policy Initiative.

RESEARCH FUNDING

2014-15 edX High School Initiative, Co-I (PI: Jeneen Graham, \$45,000)

2011-14 National Science Foundation Graduate Research Fellowship. STEM Education, *Training Elementary Math Students to Effectively Monitor Learning* (DGE-0808392). (\$132,000).

2012 University of California, Irvine, Newkirk Center for Science & Society. Graduate Student Research Fellowship, *Design and Evaluation of Brain Boost*. (Honorable Mention: \$1,000).

HONORS AND AWARDS

2014 UCI School of Education Michael E. Martinez Prize

2013 University of California, Irvine Public Impact Distinguished Fellow

2013 AERA Division C Doctoral Student Seminar

- 2012** APA Division 15 Doctoral Student Seminar
2011 UC Educational Evaluation Center Fellow
2011 AERA Motivation Special Interest Group travel award
2011 AERA Division E Graduate Student Seminar
2009 University of California, Irvine Graduate Dean Recruitment Fellowship

UNIVERSITY TEACHING EXPERIENCE

Lab Instructor, EDUC 265, Applied Regression Analysis for Education & Social Research (PhD Course) School of Education, University of California, Irvine. 2012 and 2013.
Taught STATA lab using OLS, logistic regressions, and econometric methods. Professor: Greg Duncan

Instructor, Research and Writing Workshop, in association with Boston University School of Law, Boston MA. 2002
Designed and taught 10 week writing workshop to international graduate students.

K-12 TEACHING EXPERIENCE

Teacher, Pull-Out Class, Corporate Academy North, Miami, FL. 2000
Mathematics and English instruction to prepare high school students for high school competency test.

Teacher, Computers and Art, Academy of Austin Charter School, Austin, TX. 1999
Designed and taught computer and art classes to elementary students in new charter school.

Substitute Teacher, Dade County Public Schools, Miami, FL. 1998
Short and long-term substitute teaching in high school.

Student Teacher, Fifth Grade Alternative Education, Snapper Creek Elem., Miami, FL. 1997

Swimming Instructor, Neva King Cooper Special Education Center for profoundly intellectually disabled, Dade County Public Schools, Homestead, FL. 1995-1996.

PROFESSIONAL EXPERIENCE

Education Researcher, MIND Research Institute, 2013-2014
Design, development, and study of digital games to support learning of mathematics.

Statistical Analysis & Survey Consultant, University of California, Irvine, Medical School, 2012-13
Analyzed and reported on results from an experimental medical education program.

Assessment Consultant, American Bar Association, Council on Legal Education Opportunity, 2011-12
Designed and administered assessments to evaluate program based on motivation and self-regulated learning constructs. Assisted with creation of follow-up intervention.

Educational Consultant, Seven Hills Charter School, Worcester, MA, 2006
Advised administrators and assisted in preparation for the school's state Coordinated Program Review on issues of special education, English language learners and civil rights.

Public School Liaison/Education Specialist, Massachusetts Dept. of Education, Malden, MA, 2004-06

Led and participated on teams of staff from up to six DOE units in compliance monitoring through the Coordinated Program Review System. Interviewed administrators, teachers, specialists and staff and reviewed student records and other data to rate school district compliance on areas of regulation. Interpreted education regulations and analyzed school practices during monitoring and in responding to parent complaints and information requests regarding special education and civil rights.

Judicial Law Clerk, Massachusetts Appeals Court, Boston, MA. Judge Mark Green, 2003-04
Advised Appellate judge on issues relevant to the current docket. Performed research and wrote memoranda of law and draft opinions.

NATIONAL AND INTERNATIONAL SERVICE

Ad Hoc Reviewer

Journal of Research on Educational Effectiveness, 2013-2014

Early Education and Development, 2014

Sociological Perspectives, 2014

American Educational Research Journal, 2013

Journal of Advanced Academics, 2013

Learning and Individual Differences, 2012

American Journal of Evaluation, 2012

Conference Proposal Reviewer

American Educational Research Association Annual Meeting, Division C; Studying & Self-Regulated Learning Special Interest Group, 2013; 2014; 2015; American Psychological Association Annual Meeting, *Division 15*, 2013

Webmaster & Graduate Student Committee Chair, American Educational Research Association, Studying & Self-Regulated Learning Special Interest Group, 2012-present

Division C Graduate Student Campus Liaison, American Educational Research Association, 2009-2014

UNIVERSITY SERVICE

Learning & Cognition Faculty Search Committee Student Representative, University of California, Irvine, School of Education, 2012

Math/Science Faculty Search Committee Student Representative, University of California, Irvine, School of Education, 2012

PhD Steering Committee Student Representative, University of California, Irvine, School of Education, 2010-2012

COMMUNITY SERVICE

Performer and Volunteer, Musical Theatre Village, Community Theater, Irvine, CA. 2011-present

School Site Council Member, University Park Elementary, Irvine, CA. 2011-2013

Founder and Director, Portraits for Progress, Ltd. at <http://portraitsforprogress.org>, 2007-2009
Ran an annual portrait charity event for breast cancer research. Raised and donated over \$1,000 each year.

Volunteer Coach of the Boston University Synchronized Swimming Team, Boston University, Boston, MA. 2001-2003

PROFESSIONAL AFFILIATIONS

UCI Center for Research on Cognition and Learning (CRCL)

American Educational Research Association

American Psychological Association: Division 15 & APAGS

Society for Research in Child Development

RESEARCH AND STATISTICS SKILLS

Experimental design, large-scale education field research, field administration of cognitive & achievement tests, survey and assessment construction, analysis of large datasets

Training and experience with functional Magnetic Resonance Imaging (fMRI)

OLS and logistic regressions, Multilevel Modeling, Structural Equation Modeling, Propensity Score Matching, Econometric Techniques, Social Network Analysis

Software: SPSS, Stata, AMOS, E-Prime, Mplus, FileMaker Pro, Atlas.ti, Adobe Creative Suite, UCInet

ABSTRACT OF THE DISSERTATION

Calibration of Confidence Judgments in Elementary Mathematics:
Measurement, Development, and Improvement

By

Teomara Rutherford

Doctor of Philosophy in Education

University of California, Irvine, 2014

Professor George Farkas, Chair

Self-regulated learning (SRL), the ability to set goals and monitor and control progress toward these goals, is an important part of a positive mathematical disposition. Within SRL, accurate metacognitive monitoring is necessary to drive control processes. Students who display this accuracy are said to be *calibrated*, and although calibration is a growing area of research within Educational Psychology, unanswered questions remain about the nature of calibration: how it should be measured, its role as a dynamic aspect of metacognition, and how best to improve it. This dissertation uses a rich source of data on student calibration and achievement within an online mathematics curriculum (ST Math) to approach these questions and present results on calibration as representative of a complex system of metacognition.

This dissertation presents evidence that calibration is best represented as two separate monitoring processes, one for confidence and one for uncertainty; these processes can be operationalized through the measures of Sensitivity and Specificity. In Study 1, comparisons with other commonly used measures of calibration indicate that Sensitivity and Specificity have a relative robustness to most patterns of missing data and greater strength as predictors. Other

commonly used calibration measures suffer greatly from missing data inherent in real-world patterns of question answering.

Study 2 characterizes metacognitive monitoring as part of a dynamic system that varies depending on task. In this study, variance in calibration is associated with variance in performance gain within the same student across ST Math quizzes. Both Sensitivity and Specificity are predictors of this gain, but greater confidence when correct (Sensitivity) is more strongly associated with performance gains between quizzes than is greater uncertainty when incorrect (Specificity).

Study 3 evaluates the potential of ST Math as a calibration intervention. After a year's practice and feedback with ST Math, students display greater Specificity, but lower Sensitivity, indicating that ST Math made the students more uncertain. Study 3 also explores how change in calibration is related to change in achievement, finding no relation between growth in calibration and growth in achievement, either within or outside of ST Math.

CHAPTER 1

Introduction and Literature Review

Background

Innovative math thinkers, crucial for a modern STEM workforce, are those who take ownership of their learning. This ownership can be characterized as self-regulated learning (SRL): the ability to set goals, monitor progress toward these goals, and make adjustments when necessary to ensure achievement (Zimmerman, 2008). Fostering SRL is especially important in mathematics, where students who are unable to monitor their understanding inevitably miss foundational material needed to understand more advanced concepts.

Using SRL, students incorporate feedback from prior successes and failures into the formation of strategies for future accomplishments (Greene & Azevedo, 2007; Winne, 2004). Students who are able to realistically assess their likelihood of success on a given task, and who are able to accurately reflect on previous performance, are more able to set challenging yet attainable goals, maintain motivation towards achieving these goals, and make use of strategies necessary for their success (Greene & Azevedo, 2007; Stone, 2000; Winne, 2004). Models for SRL are complex, and in order to effectively regulate their learning, students must make accurate assessments of their own capabilities and knowledge at multiple steps of the process (Greene & Azevedo, 2007). This dissertation addresses the accuracy of these assessments as *calibration* of confidence, and sets out a course of research on the measurement of calibration, its links with mathematics performance, and its malleability.

The Studies

This dissertation combines three studies centered on calibration by elementary school students addressing the research questions: (1) How can calibration best be measured? (2) Is

initial calibration accuracy linked to learning gains? and (3) Can calibration be improved?

Data from a calibration training program within the MIND Research Institute's ST Math program are utilized within all three studies. In the first, the measurement of calibration is explored to determine which of the most common measures of calibration are appropriate for real-world data of the type gathered within ST Math. In the second study, these data are analyzed to determine whether calibration is associated with greater student performance gains from pre to post-test. The third study determines whether the practice with and feedback on calibration judgments within ST Math improves calibration accuracy, comparing a group of students who have had one year of this practice with those who are just beginning to use the program.

Significance

America's ability to compete in a global market depends heavily on education (NAP, 2010). Science and math education in particular are vital for success, yet the United States public education system is not adequately developing the intellectual readiness needed to sustain the nation's economy or solve the scientific and mathematical challenges of the future (NAS, 2007; NSF, 2010). International comparisons of mathematics performance show that U.S. students fall behind other top industrialized nations (PISA, 2009), and within the U.S., achievement gaps persist between African American or Hispanic students and White or Asian American students (Fryer & Levitt, 2004; NCES, 2005; Reardon & Galindo, 2009).

Teaching students math concepts and procedures may not be enough for their success. So-called "non-cognitive" skills are also necessary to ensure lasting learning and engagement in math and other subjects (see Cunha & Heckman, 2006; NMAP, 2008). Strength in these skills can provide students with a sensitivity to know when it is appropriate to apply certain mathematical knowledge and an inclination to do so (DeCorte, Verschaffel, & Op'T Eynde,

2000). As one of these "non-cognitive" skills, SRL has been related to mathematics achievement (e.g., Fuchs et al., 2003; Schunk, 1996), as has calibration—students who are more accurate in their self-assessments achieve more (Pajares & Kranzler, 1995; Rinne & Mazzocco, 2014; Stone, 2000). Given these relations, improving the components of SRL, including calibration, may lead to improved mathematical skills, ultimately supporting achievement and engagement in mathematics, potentially far beyond the reach of the current class environment.

In order to plan interventions to improve calibration and to measure their effects, it is necessary to better understand the relation between calibration and achievement. Prior work has consistently found associations indicating that students who are better calibrated are also higher achievers (e.g., e.g., Bol, Riggs, Hacker, Dickerson & Nunnery, 2010; Chen, 2002; Tobias & Everson, 1998), often relating calibration on one task or group of tasks with an unrelated measure of performance (e.g., Jonsson & Allwood, 2003; Pallier et al., 2002). This makes it difficult to understand the function of calibration within the ecology of SRL: the link between student calibration on knowledge recall questions and performance measured as GPA may indicate that disposition toward metacognitive monitoring is linked with higher achievement, but it does not indicate whether this monitoring may serve to enhance regulation of learning from task to task or within the same task. Studying the *same* student across related tasks as is done in this dissertation can better illuminate the dynamic nature of calibration.

Inherent in the understanding of calibration and its relation with achievement is knowledge about the measurement of calibration. Recommendations as to the best way to represent a student's knowledge of what they know and don't know have previously been based on factor analysis of simulated data (e.g., Schraw, Kuch, & Gutierrez, 2013), comparisons of the biases inherent in certain indices of calibration (e.g., Masson & Rotello, 2009; Nietfeld, Enders,

& Schraw, 2006), or examinations of differential correlates of calibration accuracy depending on measure (e.g., Boekaerts & Rozendaal, 2010). Much of this work has been done using simulated or unrealistic data and has not considered the practical concerns of using each measure with data obtained from young students. Additionally, comparative studies of the predictive validity of calculation options have not been undertaken (see Schraw, 2009). Within this dissertation, real-world data obtained from meaningful educational interactions of young students with digital mathematics content provides a unique testing ground for these comparisons.

Interventions focused on calibration have largely been in non-math domains, with college-aged students, or for short durations. All three aspects may contribute to the typically weak improvements seen and the lack of transfer to achievement (e.g., Bol & Hacker, 2001; Bol, Hacker, O'Shea, & Allen, 2005; Huff & Nietfeld, 2009). Prior calibration trainings have been conducted largely in knowledge-based college classrooms (e.g., Educational Psychology). These types of domains may especially suffer from a piecemeal approach to learning antithetical to the transfer of skills (see Alexander & Murphy, 1999). In contrast, the hierarchical structure of mathematics means that it is typically taught as a *developmental progression*, where new math skills are built on previously mastered skills to form a trajectory of increasingly sophisticated thinking (see Clements & Sarama, 2009). Within this system of learning, it may be easier for students to see links between units of instruction, supporting the transfer of calibration between math topics. Just as math may be more suited to interventions focused on calibration improvement, there may be ages at which students may benefit more from this type of training. Middle childhood, as metacognitive skills begin to emerge, may be an ideal time to bolster SRL, including calibration (Cunha & Heckman, 2006; Davis-Kean, Jager, & Collins, 2009). As with any skill, calibration of metacognitive judgments requires extensive opportunity for practice to

become a transferable habit of mind (see Alexander & Murphy, 1999). Prior trainings have involved only a handful of practice opportunities over periods lasting at most a semester (e.g., Nietfeld, Cao, & Osbourne, 2006). Study three within this dissertation focuses on an intervention to improve calibration of metacognitive judgments in math, with students in middle childhood, and over the course of one year with multiple practice opportunities (30+). The domain, age, and scope of this study will be a unique contribution to the calibration intervention literature.

These three studies combine research on measurement of calibration, its relation with achievement, and the malleability of calibration in response to an intervention. This dissertation will contribute to the body of literature in these areas and also inform future interventions for improving calibration of metacognitive judgments within math and beyond, potentially having a significant effect on SRL and achievement, with further potential to improve these skills beyond the immediate subject or class.

Structure

The following section contains a discussion of the conceptual framework for this dissertation and a description of self-regulated learning and the role of calibration within. Each study within the dissertation will then be laid out in turn: the study-specific literature will be reviewed, the questions presented, and the methods and results described. These results will be discussed within each study's chapter and tied together in a final overarching conclusion.

Literature Review

Theoretical Framework, Social Cognitive Theory and Metacognition

The SRL cycle is situated within a *social cognitive* perspective—one that seeks to explain individual processes in interaction with the environment (Schunk, Pintrich, & Meece, 2008; Zimmerman, 1989). Within this perspective, human functioning is explained by the interplay of personal factors, environmental factors, and behaviors (Bandura, 1986). Activities within the SRL cycle mediate the relations between student and environment to culminate in learning and achievement (Pintrich, 2004). Personal characteristics of the learner and characteristics of the context each contribute to the student's regulation of learning. The context influences student motivation and behavior within the SRL cycle and it also must provide opportunity for the learner to engage in SRL: there must exist the potential for the learner to have control over his/her cognition, motivation, environment, and/or behavior (Pintrich, 2004).

SRL is also informed by theories of metacognition, known colloquially as “thinking about thinking.” Drawing on ideas from the cognitive revolution and the turn away from behaviorism, psychology within the 1960s and 1970s focused on processes that occurred within the individual (Dunlosky & Metcalfe, 2009). These ideas about cognition, along with Piagetian theories of development, were brought together in Flavell’s (1979) introduction of metacognition. Flavell defined metacognition as knowledge and cognition about cognitive phenomena, and described the elements of metacognitive knowledge (e.g., knowledge of yourself as a person, knowledge about tasks, and knowledge about strategies) and metacognitive experiences: the conscious consideration of the task at hand, such as a “momentary sense of puzzlement” (1979, p. 908). Flavell’s ideas can be contrasted with the modern study of metacognition within a social cognitive framework in the perception of “hot” vs. “cool”

cognition. Flavell described metacognitive experiences as most likely to occur in moments of cognitive purity “where high affective arousal or other inhibitors of reflective thinking are absent” (p. 908). Real-world metacognitive experiences, such as senses of puzzlement, exist within a context of goals and emotions (see Bandura, 2001) and interact with features of the environment and the agent’s perception of those features (Efklides, 2011; Zimmerman, 2000). Because aspects of the current study investigate self-regulated learning within the context of classrooms, subjects, and learning environments, and emphasize the agentic nature of the student participants, social cognitive theory is a more fitting theoretical framework than traditional metacognition. However, studies on SRL and calibration within a social cognitive framework share a history with studies of calibration and metamemory within a more purely metacognitive framework, and therefore metacognitive research is discussed within the literature review and considered as informative to the study.

Self-Regulated Learning and Metacognitive Judgments

Emerging in the 1980s, the concept of SRL was one that drew on information processing, social cognitive theory, and the ideas inherent in the earlier lines of research on self-regulation and metacognition (Dinsmore, Alexander, & Loughlin, 2008; Zimmerman, 2001). Within a social cognitive perspective, SRL is often represented by the three stages seen in Figure 1.1. Learners set goals, monitor their progress as they perform actions in pursuit of these goals, and evaluate their performance in light of these goals to make adjustments to their goals or strategies when reentering the cycle (Pintrich2000; Zimmerman, 1989; Zimmerman, 2008; Zimmerman, 2000). Although the model presented in Figure 1.1 is the conceptualization of SRL born from the social cognitive perspective, models of SRL from other perspectives share similar features, namely purposeful use of strategies or processes, a self-oriented feedback loop, and a motivation

to engage in the learning process (Zimmerman, 1998, 2001).

{Insert Figure 1.1}

Accurate metacognitive judgments are important at each stage of the process. In the planning or forethought phase, students rely on their self-efficacy or judgment that they will be able "to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986, p. 391). Goals can be adjusted in light of this self-efficacy judgment. An accurate judgment would result in a goal that is attainable; setting too lofty a goal might result in failure accompanied by discouragement and disengagement. The ideal self-efficacy is one that is slightly positively biased, allowing for appropriate goal-setting, and also for persistence in the face of obstacles (Bandura, 1986; Schunk et al., 2008; Winne, 2004). This persistence can be seen in the performance phase as learners must maintain their sense of their goals despite challenges they encounter. Also within the performance phase, students adjust their strategies and resource allocation as they monitor their success with relation to their goals and sub-goals (Nelson, 1996; Pintrich, 2004; Winne, 2001). In the final phase, as students evaluate their ultimate goal attainment and the usefulness of their strategies to obtain this goal, accurate metacognitive judgments will guide students to make necessary changes.

The process of studying for a test is an oft-used example of SRL (e.g., Nelson, 1996). Within this example, a student may set the goal that she would like to earn an A on her Educational Psychology midterm. To set this goal she uses information about herself as a learner generally and in this domain, information about the task (in this case the test), and information about what strategies are likely to yield the best results (see Efklides, 2011; Flavell, 1979). This information is used to determine whether an A is an attainable goal, and what actions need to be taken to attain this goal. An error in her metacognitive knowledge or her metacognitive

judgments regarding the accuracy of this knowledge will result in an error in goal-setting and/or planning.

Our student has set a goal and has determined that the best course of action is to study using the instructor-provided outline and practice test—it worked well for her on the last quiz she took in this same class. As she studies, she monitors her understanding: she reads a question and asks herself whether she knew the answer, she speaks the answer to herself and has a sense as to whether it is the correct one. As she does this she also asks herself about the material in general—making determinations as to which topic she knows and does not know and dedicating extra time where she thinks she has holes in her understanding. An error in these judgments might cause her to misallocate study time. As she assesses her learning she evaluates how well her strategies are working. She identifies that she does not understand as much as she would like and changes strategy, perhaps by seeking help from the TA during office hours.

At each stage of the process, the student's successful regulation of her learning is contingent on making accurate metacognitive judgments. At each stage is also the opportunity for her to enhance her SRL with interaction with the environment (see Zimmerman, 2000). She uses resources provided by the instructor to scaffold her self-regulation, she uses the self-quiz technique because it was modeled for her and she practiced it, and she seeks help because she is aware of the resource the TA provides and has been encouraged to use it. In this way, her environment supports current self-regulation and development of future SRL skills.

Accuracy of Metacognitive Judgments: Calibration

For a given task, student calibration accuracy is defined by comparing student predictions or postdictions with actual achievement. Predictions are those judgments made before undertaking a task, and postdictions are those made after completing a task. For example, a

student about to take a math test can make a *prediction* as to how many questions he or she will answer correctly and then, after attempting the questions, the student makes a *postdiction* as to how many he or she did answer correctly. Two types of calibration *bias* are possible. When a student predicts or postdicts accuracy above that which he actually attains, that student is overconfident. This is in contrast to underconfidence, when a student's indication of his likelihood of success is below his attainment of actual success.

Differences between perception of performance and actual or objective measures of performance are well documented in the literature—people in general tend to be overconfident (Chen, 2002; Kruger & Dunning, 1999; Stone, 2000). Individual differences may influence relative overconfidence. Those who have limited knowledge about a domain tend to be more overconfident (Kruger & Dunning, 1999). Stages of domain learning as described in Alexander and Murphy (1999) may help to illuminate some of these differences. Alexander and Murphy describe new learners within a domain as those who must rely heavily on general metacognitive strategies, yet do not have the domain-specific knowledge to use them efficiently. Nor do they have the cognitive resources free to focus attention on metacognitive pursuits (see Alexander & Murphy, 1999; Avery & Smillie, 2012). Lack of knowledge within a domain may also result in individuals using the wrong resources with which to base their judgments. Dinsmore and Parkinson (2013) found that students who were more poorly calibrated often relied on multiple factors when making metacognitive judgments, perhaps reflecting their failure to understand the most important cues within the test information. Kruger and Dunning (1999) note this link between knowledge and calibration as a "dual burden"—individuals lack both the knowledge of material and the knowledge that they don't have this knowledge (p. 1,121), handicapping their motivation to acquire the knowledge they need.

Generally, higher achieving students are found to be better calibrated, and have been found by some researchers to exhibit underestimation bias, if any (Stone, 2000). This has been replicated across domains and age groups. With elementary students, the relation between accurate calibration and achievement has been found in math (e.g., Barnett & Hixon, 1997; Tobias & Everson, 1998), reading (e.g., Fajar, Santos, & Tobias, 1996; Romero & Tobias, 1996), social studies and spelling (e.g., Barnett & Hixon, 1997), and in playing computer games (e.g., Nietfeld, Minogue, Spires, & Lester, 2013). This calibration/achievement link was also found in middle/high school math students (e.g., Bol, Riggs, Hacker, Dickerson & Nunnery, 2010; Chen, 2002; Chen, 2007, in an international sample; Pajares & Kranzler, 1995), and in undergraduate students within knowledge-based courses like Research Methods (e.g., Bol & Hacker, 2001; Bol et al., 2005). Many of these studies use researcher-created measures of achievement; little work in the K-12 arena has used real graded assignments or relevant achievement measures (Hacker, Bol, & Keener, 2008). These associations between calibration and achievement are explored within the studies of this dissertation. The following chapter discusses issues inherent in the calculation of measures of calibration and compares ten commonly used measures of calibration—investigating both their practicality and their predictive validity.

References

- Alexander, P. A., & Murphy, P. K. (1999). Nurturing the seeds of transfer: a domain-specific perspective. *International Journal of Educational Research*, 31(7), 561–576.
doi:10.1016/S0883-0355(99)00024-5
- Avery, R. E., & Smillie, L. D. (2012). The impact of achievement goal states on working memory. *Motivation and Emotion*. doi:10.1007/s11031-012-9287-4
- Bandura, A. (2001). Social cognitive theory: an agentic perspective. *Annual review of psychology*, 52, 1–26. doi:10.1146/annurev.psych.52.1.1
- Bandura, Albert. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs N.J.: Prentice-Hall.
- Barnett, J. E., & Hixon, J. E. (1997). Effects of Grade Level and Subject on Student Test Score Predictions. *The Journal of Educational Research*, 90(3), 170–174. doi:10.2307/27542087
- 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(5), 372–382. doi:10.1016/j.learninstruc.2009.03.002
- 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(2), 133–151. doi:10.2307/20152656
- 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(4), 269–290. doi:10.2307/20157403
- Bol, L., Riggs, R., Hacker, D. J., Dickerson, D.L., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education*, 21, 81-96.

- Chen, P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77–90.
- Chen, Peggy. (2007). A Cross-National Comparison Study on the Accuracy of Self-Efficacy Beliefs of Middle-School Mathematics Students. *Journal of Experimental Education, 75*(3), 221–244.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: the learning trajectories approach*. Taylor & Francis.
- Cunha, F., & Heckman, J. J. (2008). Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources, 43*(4), 738–782. doi:10.3368/jhr.43.4.738
- Davis-Kean, P. E., Jager, J., & Andrew Collins, W. (2009). The Self in Action: An Emerging Link Between Self-Beliefs and Behaviors in Middle Childhood. *Child Development Perspectives, 3*(3), 184–188. doi:10.1111/j.1750-8606.2009.00104.x
- De Corte, E., Verschaffel, L., & Op 't Eynde, P. (2005). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation*. Elsevier.
- Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the Conceptual Lens on Metacognition, Self-regulation, and Self-regulated Learning. *Educational Psychology Review, 20*(4), 391–409. doi:10.1007/s10648-008-9083-6
- Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. doi:10.1016/j.learninstruc.2012.06.001
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE.

- Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist, 46*(1), 6–25.
doi:10.1080/00461520.2011.538645
- Fajar, L., Santos, K., & Tobias, S. (1996, October). *Knowledge monitoring among bilingual students*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. doi:10.1037/0003-066X.34.10.906
- Fryer, R., & Levitt, S. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics, 86*(2), 447-464.
doi:10.1162/003465304323031049
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade student’ mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology, 95*(2), 306–315.
doi:http://dx.doi.org/10.1037/0022-0663.95.2.306
- Greene, J. A., & Azevedo, R. (2007). A Theoretical Review of Winne and Hadwin’s Model of Self-Regulated Learning: New Perspectives and Directions. *Review of Educational Research, 77*(3), 334–372. doi:10.3102/003465430303953
- Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 429-455). New York: Psychology Press

- Huff, J., & Nietfield, J. (2009). Using Strategy Instruction and Confidence Judgments to Improve Metacognitive Monitoring. *Metacognition and Learning*, 4(2), 161–176.
- Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences*, 34(4), 559–574. doi:10.1016/S0191-8869(02)00028-4
- 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(6), 1121–1134.
- Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 35(2), 509–527. doi:10.1037/a0014876
- National Academies Press (2010). *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5*. Retrieved from http://books.nap.edu/catalog.php?record_id=12999
- National Academies of Science, (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. The National Academies Press, Washington, DC:1-13. Retrieved from <http://www.nap.edu>
- National Center for Education Statistics (2005). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective*. Retrieved from <http://nces.ed.gov/pubs2005/2005003.pdf>

- National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Retrieved from <http://www.ed.gov/about/bdscomm/list/mathpanel/index.html>
- National Science Foundation (2010). *Preparing the Next Generation of STEM Innovators: Identifying and Developing our Nation's Human Capital*. Retrieved from <http://www.nsf.gov/nsb/publications/2010/nsb1033.pdf>
- Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102–116. doi:<http://dx.doi.org/10.1037/0003-066X.51.2.102>
- 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*(2), 159–179. doi:10.1007/s10409-006-9595-6
- Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo Comparison of Measures of Relative and Absolute Monitoring Accuracy. *Educational and Psychological Measurement*, *66*(2), 258–271. doi:10.1177/0013164404273945
- Nietfeld, J. L., Minogue, J., Spires, H. A., & Lester, J. (2013, April). Girls and games: Examining the performance and self-regulation of girls in a science gaming environment. Paper presented at the annual meeting of American Educational Research Association, San Francisco, CA.
- Pajares, F., & Kranzler, J. (1995). Self-Efficacy Beliefs and General Mental Ability in Mathematical Problem-Solving. *Contemporary Educational Psychology*, *20*(4), 426–443. doi:10.1006/ceps.1995.1029

- Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The Role of Individual Differences in the Accuracy of Confidence Judgments. *The Journal of General Psychology, 129*(3), 257–299. doi:10.1080/00221300209602099
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 451–502). Elsevier.
- Pintrich, P. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review, 16*(4), 385–407. doi:10.1007/s10648-004-0006-x
- Programme for International Student Assessment (2009). *Science competencies for tomorrow's world*. Organisation for Economic Co-operation and Development. Retrieved from www.pisa.oecd.org
- Reardon, S. F., & Galindo, C. (2009). The Hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*(3), 853 -891. doi:10.3102/0002831209333184
- Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing Right From Wrong In Mental Arithmetic Judgments: Calibration Of Confidence Predicts The Development Of Accuracy. *PLoS ONE, 9*(7), e98663. doi:10.1371/journal.pone.0098663
- Romero, R. & Tobias, S. (1996, October). *Knowledge monitoring and strategic study*. Paper presented at a symposium on "Metacognitive Knowledge Monitoring" at the annual convention of the Northeastern Educational Research Association, Ellenville, NY.
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. doi:10.1007/s11409-008-9031-3

- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57.
doi:10.1016/j.learninstruc.2012.08.007
- Schunk, D. H. (1996). Goal and Self-Evaluative Influences During Children's Cognitive Skill Learning. *American Educational Research Journal, 33*(2), 359–382.
doi:10.3102/00028312033002359
- Schunk, D. H., Pintrich, P. R., & Meece, J. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Prentice Hall.
- Stone, N. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. doi:10.1023/A:1009084430926
- Tobias, S. & Everson, H. T. (1998, April). *Research on the assessment of metacognitive knowledge monitoring*. Paper presented at the annual convention of the American Educational Research Association, San Diego.
- Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-Regulated Learning and Academic Achievement*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.
- Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466–488.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. doi:http://dx.doi.org/10.1037/0022-0663.81.3.329

- Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D.H. Schunk & B.J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 1-19). New York: Guilford Press.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–40).
- Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.) (pp. 1–38). Mahwah, NJ: Erlbaum.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183. doi:10.3102/0002831207312909

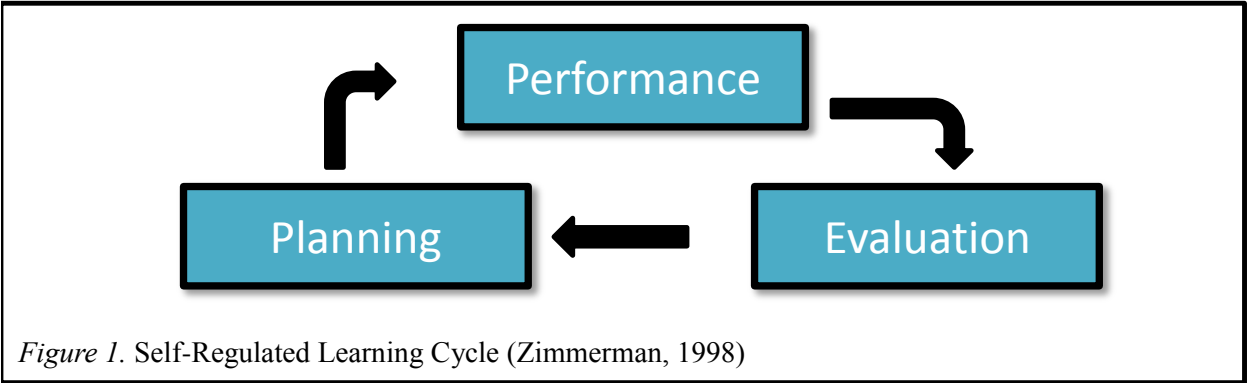


Figure 1. Self-Regulated Learning Cycle (Zimmerman, 1998)

CHAPTER 2

Study 1: The Measurement of Calibration in Real Contexts

Calibration, although generally referring to knowledge of what one knows and doesn't know, can be operationalized in a variety of manners. When operationalizing calibration, researchers focus on three dimensions: timing, grain size, and calculation of measurement. Timing focuses on when metacognitive judgments are elicited, before or after an event, described above as predictions and postdictions. Timing has important indications for conducting research on calibration—for example, postdictions are generally more accurate than predictions (Hacker, Bol, & Keener, 2008). Predictions and postdictions may implicate different levels of metacognitive knowledge within an individual. In a testing context, a prediction before a student attempts to answer an item is completed with imperfect information about the actual material to be tested and may reflect a more general sense of self-efficacy about the domain of assessment. In contrast, postdictions are completed after experiencing attempt at recall or problem-solving and are therefore more closely tied to the material tested.

Grain size refers to the level at which the judgments are elicited (see Schraw, 2009). A student could be asked for a pre or postdiction for how well they think they will do/did in a course, on a test, or on an individual item. Studies within Educational Psychology often look at macro levels of calibration, for example, comparing actual percent of items correct with a student's pre or postdiction of items correct (see Keren, 1991; e.g., Barnett & Hixon, 1997; Bol & Hacker, 2001). As seen in Figure 2.1, this can ignore important information. Imagine that two students, Sarah and Jenny, each took a five-question quiz and gave item-by-item confidence ratings (confident/not confident). Looking only at the macro level, Jenny appears perfectly calibrated—her level of confidence matches her level of accuracy; however, she is only properly

calibrated on 60% of the items. In comparison to macro-level analysis, item-by-item analysis allows a more detailed, and likely more accurate, view of the process of forming metacognitive judgments.

Finally, choice of calculation of calibration affects conclusions drawn. Researchers can focus on absolute calibration (e.g., Bol & Hacker, 2001; Huff & Nietfeld, 2009) or can investigate the direction of the calibration (e.g., Chen, 2002; Mengelkamp & Bannert, 2010). In Figure 2.1, both Sarah and Jenny have the same level of calibration (looking item-by-item), but Sarah displays an overconfident bias whereas Jenny is not biased in either direction. It is this choice of how to calculate the level of calibration when using item-by-item comparisons upon which this study focuses, asking (1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments? and (2) Among these measures, which display the greatest predictive validity?

{Insert Figure 2.1}

Comparisons of Calibration Measures

In selecting measures of calibration, prior research has noted the importance of aligning the purpose of the study with the selected measure (see Boekaerts & Rozendaal, 2010; Nietfeld, Enders, & Schraw, 2006; Schraw, 2009). Various measures may be complimentary in that they can provide information on absolute accuracy, bias, or also the ability to distinguish between correct and incorrect items—each may be more or less useful in light of particular research questions (see Boekaerts & Rozendaal, 2010; Schraw, 2009). However, practical considerations beyond the match with research question may also guide the choice of method. The balance between sensitivity and ease of use has been one such issue. It has been suggested that for young children especially, measures with fewer choices reduce the cognitive load and allow for more

accurate calibration scores (see Huff & Nietfeld, 2009; see e.g, Lyons & Ghetti, 2011). As seen with the example of Sarah and Jenny (Figure 2.1), students can indicate on a dichotomous measure whether they feel confident or not confident.

{Insert Figure 2.2}

The use of such a dichotomous measure in relating accuracy to judgments of confidence results in a 2x2 contingency table with cells described as in Figure 2.2. Looking to our examples: of the five quiz questions, Sarah would have one question in cell A, two each in cells B and D, and none in cell C. Jenny would have two questions in cell A and one each in the other three cells. Numerous indices have been created for the calculation of agreement based on the contents of these cells (see Feuerman & Miller, 2008; Schraw et al., 2013). Table 2.1 presents a number of common indices expressed as functions of cells A through D and largely draws on descriptions of these formulas as presented in Schraw and colleague's (2013) work. Some measures have emerged as more popular than others: Gamma (e.g., Mengelkamp & Bannert, 2010; Thiede, Anderson, & Therriault, 2003), d' or discrimination (e.g., Boekaerts & Rozendaal, 2010; Macmillan & Creelman, 1996), and G Index (e.g., Schraw, 1995; Tobias & Everson, 2002) have been particularly popular within metacognition research. These and other measures have theoretical justifications (e.g., Gamma may be most useful in determining consistency of judgments whereas G Index may be most useful in measuring changes in calibration, see Nietfeld et al., 2006), but there are also practical ramifications of the selection of one measure over another. Due to the nature of the formula calculations, the distribution of data within the 2x2 contingency tables affects each of the measures differently. For some, the lack of selections that fall in certain quadrants is especially problematic. As an example, Gamma is undefined

when certain combinations of quadrants are missing (A and C, A and B, D and C, D and B) and can be heavily distorted when even one quadrant is zero (see Kuch, 2012).

These distortions from missing quadrants have been quantified and discussed in prior research (e.g., Kuch, 2012; Mason & Rotello, 2009; Nietfeld et al., 2006; Schraw, Kuch, & Roberts, 2011). In particular, previous work has examined the extent of distortion due to number of test questions and difficulty of questions (e.g., Schraw et al., 2011; Kuch, 2012), the comparative distortion between measures (e.g., Kuch, 2012; Nietfeld et al., 2006), and solutions for eliminating this distortion (e.g., Hautus, 1995; Miller, 1996; Schraw et al., 2011). The bulk of this research is conducted by examining the behavior of the measures using simulated data. A typical process uses a Monte Carlo simulation to create responses for questions to tests of lengths from 6 to 1,000 questions (often assuming the correct/non-distorted estimates will be present at 1,000 questions). To simulate a distribution of responses due to chance, each question response is randomly assigned to one of the four quadrants resulting in approximately 25% of the responses in each cell. To simulate a moderately accurate condition, which is often assumed to approximate real-life conditions (see Kuch, 2012; Nietfeld et al., 2006; Schraw et al., 2013), 50% of the responses are assigned to cell A (confident and correct) and then the remaining 50% are randomly assigned across all four cells. This results in a distribution of 62.5% in cell A and 12.5% in each of the other cells. Based on such simulated datasets, Nietfeld and colleagues (2006) concluded that G Index was more reliable across varying test sizes than Gamma, a conclusion supported by Kuch (2012). Schraw and colleagues (2011) suggested that for Gamma to be reliable it needed to be calculated from moderately difficult tests of at least 20 questions. This suggestion for test size was based on data behavior that simulated an equal likelihood of being in cells B through D, even in conditions meant to approximate tests of moderate difficulty.

However, theory surrounding metacognitive judgments does not support this distribution of responses. As difficulty increases, it is more likely that test-takers will make *overconfident* judgments (see Kruger and Dunning, 1999), likely resulting in a paucity of quadrant D responses, even when test-takers make more incorrect responses.

{ Insert Table 2.1 }

To attempt to avoid distortion from zero quadrants, two tactics have been previously used. A number, such as .05, can be added either to only the missing quadrant or to all four quadrants (see Hautus, 1995). Schraw et al. (2011) and Miller (1996) demonstrated with simulated data that adding a number to the missing quadrant (most likely quadrant D in Schraw et al., 2011) did not eliminate data distortion and that distortion varied depending on the exact number added and on the value for calibration that would have been observed without distortion (based on simulated data of at least 1,000 questions). Hautus (1995) noted that although neither commonly used convention for handling missing quadrants completely replicated true non-problematic data, the practice of adding a value across all quadrants came closer. Schraw and colleagues (2011) discouraged researchers from the practice of modifying the data, noting that modifying the instrument instead (as noted above, to a moderately difficult >25 question test) should solve the problem of missing quadrants.

Schraw (2009) describes both practical and theoretical limitations to a number of measures of calibration, noting that although Monte Carlo experiments can inform researchers about certain properties of these measures, comparative studies of predictive validity were much needed and had been virtually non-existent. In a later paper, Schraw and colleagues (2013) again turned to simulated data to move beyond the practical difficulties of distortion due to zero quadrants to concentrate on the interrelation of common measures. The authors classified each of

the measures in Table 2.1 into one of five interpretive families based on the main purpose for each measure (diagnostic efficiency, agreement, association, binary distance, and discrimination) and noted that there were theoretical reasons for possible degrees of correlation between each of the measures. Using simulated datasets of 1,000 question quizzes, they conducted confirmatory factor analyses to test three competing models of metacognitive monitoring: the first, based on the Nelson and Narens (1990) one-factor solution, the second, specifying a two-factor solution with sensitivity and specificity as orthogonal processes subsuming variance in all the other measures (see Feuerman & Miller, 2008), and the third, specifying five interrelated factors based on the theoretical families. The authors concluded that the second model was the best-fitting for their simulated data in both the chance and moderate accuracy conditions and that sensitivity and specificity, together in a combined model, should be the best indicators of metacognitive monitoring.

Schraw and colleagues (2013) had theoretical reasons for supporting their advocacy of a combined sensitivity/specificity model; however, their conclusions were based off of simulated data and did not investigate the predictive validity of each of the measures studied. A comparison of these common measures of calibration has not been undertaken using non-simulated data, especially those from an authentic learning task. Most studies of calibration with authentic learning tasks include relatively small samples using tests of limited size (see Schraw et al., 2013; e.g., Huff & Nietfeld, 2009; Pajares & Miller, 1997). This makes comparisons between measures difficult. Even those that have used multiple measures (e.g., Allwood, Jonsson, & Granhag, 2005; Boekaerts & Rozendaal, 2010) do not explore practical limitations of the data and often focus on calibration as an outcome (examining correlates between score and individual and/or task), without examining differences in predictive validity between scores. The current

paper utilizes data that addresses both the shortcomings of simulated data and authentic data as typically used. The data in this dissertation are authentic in that they come from student interactions with learning materials administered as part of their normal mathematics classes and the data are not subject to the typical limitations of real-world data in that the sample size is large (over 4,000 students) and aggregated across the year's curriculum to produce a test with over 200 questions. Using these data I examine how the actual distortion from zero quadrants affects the calculation of different calibration measures and how these measures, once calculated, differentially predict measures of achievement gains.

Method

Research design. The proposed dissertation uses data from an ongoing study of ST Math funded by an IES grant to a partnership between MIND, the Orange County Department of Education, and researchers at the University of California, Irvine. Within the larger study, the effectiveness of the digital mathematics curriculum, Spatial Temporal (ST) Math, was evaluated using a randomized control trial of 52 schools (see Rutherford et al., 2014). The 52 elementary schools in the study included two cohorts with a staggered implementation design. The studies within this dissertation will concentrate on Cohort 2 schools, which began implementing ST Math in the 2009-2010 school year. Participating students played the ST Math games 90 minutes each week during each school year starting in 2009-2010.

Research population. The overall ST Math study sample consisted of all second through fifth grade students in 52 low-performing schools within ten districts in Southern California. Each school enrolled approximately 200 to 800 students in second through fifth grade in a given year. This dissertation concentrates on students within the 18 Cohort 2 schools. Descriptive statistics for the study sample are shown in Table 2.2. Of the participating students, 52% were

male. The participants were largely Hispanic (85%), with White (8%), and Asian (3%). Other ethnicities, including African American and Native American, comprised another 3% of the sample. Eighty one percent of the students were eligible for free or reduced lunch and 65% of the students were English Language Learners (ELLs). In comparison to the populations of the county and to the state of California as a whole (CADOE STAR, 2011), the study schools contained a larger percentage of Hispanic students, ELLs, and students eligible for free or reduced lunch. Looking further at student subgroups, among the ELL students in the sample, 96% were Hispanic, and among Hispanic students, 74% were ELLs. This study focuses on the 4,281 students who were using ST Math in the 2010-2011 school year: approximately half of the second, third, and fifth graders in the Cohort 2 study schools, and all of the fourth graders.

{Insert Table 2.2}

Instruments/measures/sources of data.

ST Math quiz data. Within ST Math, students completed up to 24 mathematics objectives, depending on grade level. As students started a new objective module, they took a five-question pretest on the content within that module and specified their confidence (sure or not sure) in each answer they gave (see Figure 2.3). After the module, they took a five-question posttest, also selecting their confidence level. The combination of this accuracy and confidence data provided information on student calibration. MIND provided in-game quiz scores and calibration measures for each of the students who engaged with the ST Math curriculum during the 2010-2011 school year. Data included item-by-item quiz answers, accuracy, and confidence ratings. Each year included up to 48 quizzes (considering pre and post separately), depending on grade-level, administered to students as they completed the ST Math curriculum. Calibration measures for the current analyses were calculated from scores aggregated across all quizzes

taken during the year's curriculum. This includes up to 270 questions for second grade, 280 questions for third grade, 270 questions for fourth grade, and 230 questions for fifth grade.

{Insert Figure 2.3}

Demographics. Gender, ethnicity, free/reduced lunch, and ELL status were reported by the school districts along with the CST data. Ethnicity is represented in the analyses by five groups: Hispanic, Vietnamese, Black, White, and Other Ethnicity, to represent the largest ethnic groups within the sample. Reported English Language Learner (ELL) status was determined by schools as measured by the California English Language Development Test (California Department of Education, 2011). Federal free/reduced lunch program eligibility stands in as a measure of student socioeconomic status.

Analysis

To answer the first research question, regarding which measures of calibration can accommodate real-world data, quiz question data were aggregated across all the objectives to provide for the largest possible sample of questions. After aggregating, quadrants A through D were summed to represent each student's quadrant totals for each combination of confidence and accuracy. Two types of analyses were conducted to examine the relation between zero values in one or more of the quadrants and the number of questions answered. As a first step, the complete data were used for each student and logistic regressions were calculated to predict the likelihood of a zero in each quadrant based on the number of questions answered. Because student factors related to their likelihood for both accuracy and confidence were hypothesized to also be related to student ability to complete the curriculum and therefore their number of questions answered, additional analyses were conducted using data from only those students who had completed a substantial portion of the curriculum (200 questions). Within these students, 200 randomly drawn

(without replacement) datasets of 25, 50, 75, 100, and 150 questions each were created, and from these, the percentage of students with zero quadrants was examined. Random selections of questions were chosen to control for variation in question difficulty, hypothesized to be related to both student accuracy and confidence. This analysis examined the possibility of incalculable measures due to zero quadrants by using a range of realistic quiz lengths (see Nietfeld et al., 2006; Schraw et al., 2011).

After examining the possibility of zero quadrants, each of the ten measures described in Schraw et al. (2013) was analyzed to determine the proportion of students for whom each of the measures was not calculable (due mostly to zero denominators). For this analysis, the full sample of student data with varying completion rates was utilized. The measures were then recalculated after first adding 1 to each quadrant so that no students would have zero quadrants and all measures would be calculable. Means for each measure calculated from the unaltered sample (not including those students with incalculable measures) were compared to those from the measures calculated with the quadrants modified to ensure no zero quadrants.

To answer the second research question, regarding the predictive validity of the ten calibration measures, the data were first limited to students for whom each of the ten measures was calculable. Separate regressions were conducted to examine the association between pretest calibration and posttest accuracy for each measure, controlling for pretest accuracy, student grade level, number of quizzes completed, and student demographic variables (gender, ethnicity, ELL and free/reduced lunch status, and grade-level). An additional model was examined considering sensitivity and specificity together, as recommended by Schraw et al. (2013). These analyses were replicated with the full sample of students with the measures adjusted to eliminate zero quadrants.

Results

Zero quadrants. Within the ST Math curriculum, not all students took all quizzes—teachers may have "skipped" some students past certain objectives or reordered the objectives, or, because of the self-paced nature of the program, students may not have reached the final objectives. The mean number of questions completed by students was 156.56 (SD=73.43), differing slightly between the grade levels (see Table 2.3). Figure 2.4 displays the distribution of each quadrant within the 2x2 contingency table of accuracy and confidence. Any given student, however, may have a distribution of confidence and accuracy that filled only some of these quadrants. As noted above, due to the nature of calibration calculations, a zero in any given quadrant may make measures of calibration incalculable. 796 of the 4,281 students (19%) had zeroes in at least one quadrant: less than one percent of students had zeroes in quadrants A or B, indicating that most students had at least one question on which they were confident and correct and at least one question on which they were confident, but not correct. However, 15% of students had no unconfident and correct answers (quadrant C), and 9% had no unconfident and incorrect answers (quadrant D). Six percent of students had zeroes in two quadrants, and less than one percent of students had zeroes in three quadrants.

{Insert Table 2.3}

{Insert Figure 2.4}

Looking at the relation between likelihood of having a zero in a quadrant and number of questions completed, I focused only on quadrants C and D, which both involved student determinations of uncertainty. Results from logistic regressions are presented as odds ratios and marginal effects in Table 2.4. Students who completed more questions were less likely to have zeroes in quadrants C and D. For quadrant C, at the mean of total questions completed (157

questions), completing one more question is associated with a .10 percentage point decrease in the probability of never making a judgment of uncertainty when a student has the question correct. This would indicate that a student would have to answer 150 more questions (307 total) to bring her probability of having a quadrant C zero from 15%¹ to nothing. Similarly, for quadrant D, at the mean of total questions completed, completing one more question is associated with a .10 percentage point decrease in the probability of never making a judgment of uncertainty when incorrect. A student would have to answer another 90 (247 total) questions to bring her probability of having a zero in quadrant D from 9% to nothing.

{Insert Table 2.4}

Student factors were related to the number of quiz questions completed (see Appendix A, Table 1). Statistically significant associations emerged between these factors and question completion: in all grades but fifth grade, boys completed more questions than girls and those students eligible for free/reduced lunch completed fewer questions than those who were not. In all grades, Asian students completed more questions than Hispanic students, and in all grades but third, ELLs completed fewer questions than those who were not labeled as such. Additionally, variation in the difficulty of the questions may have affected both the proportion of students who answered the question correctly, and the students' judgments of confidence. Depending on grade-level, the average question accuracy was between 60 and 66%, with standard deviations around 20%.

To explore the association between number of questions completed and likelihood of having a zero in one of the quadrants without the confounding factor of student progress, the data were limited to those of students who had completed at least 200 questions. Demographic

¹ 15% is the mean number of quadrant C zeroes using the entire dataset with a question range from 5-280, depending on grade—not necessarily the number of quadrant C zeroes at the mean question number, 157. Likewise, the mean number of quadrant D zeroes using the entire dataset is 9%.

information on this reduced sample of 1,341 students is provided on the left side of Table 2.5. This sample represented 31% of the original sample. This percentage varied across grades: 37% of second graders, 29% of third graders, 30% of fourth graders, and 22% of fifth graders. Tables 2-5 in Appendix A present the results of from the 200 randomly drawn datasets of 25, 50, 75, 100, and 150 questions. The mean percentage of students missing values from each quadrant is presented as is the 98% confidence intervals around these means. The mean number of zero quadrants is also presented for the total number of questions (M=220-258, depending on grade). For each quadrant, the more questions that are used, the less likely there was a zero value in that quadrant. Using the mean number of zeros for each of the randomly drawn datasets, at 50 questions, for second, fourth, and fifth graders, there were no students without at least one question on which they were both correct and confident (quadrant A). Third graders continued to have a few students missing data in this quadrant until 150 questions. Quadrant B, those questions which students got incorrect, but indicated confidence, followed a similar pattern. At 25 questions, between 3.69% and 8.17% of students, depending on grade, did not have any answers that fell in this quadrant. This number dropped to between .01% and 4.87% at 150 questions. Quadrants C and D, the quadrants representing student judgments of uncertainty, were missing from a large proportion of students. Approximately 40% of students never made judgments of uncertainty when they had the correct answer (quadrant C) in a randomly drawn sample of 25 questions. At 150 questions, this number dropped to between 9.03% and 16.59%, depending on grade. When samples of 25 questions were randomly drawn, approximately 30% of students never made judgments of uncertainty when they had the incorrect answer (quadrant D). At 150 questions, this dropped to between 5.07% and 15.02%, depending on grade.

{Insert Table 2.5}

{Insert Table 2.6}

Measures of calibration. Table 2.6 presents the descriptive statistics from the calculation of the ten measures of calibration described in Schraw et al. (2013) calculated with the full sample of 4,281 students. As predicted from the presence of zero quadrants within these data, not all measures could be calculated for all students. Sensitivity, Specificity, Simple Match, G Index, and Sokal Reverse could be calculated for almost all the students—98% or more of the students in the sample had valid data for these measures. Odds Ratio, Gamma, and Phi suffered moderately from the presence of zero quadrants. For example, in fourth grade, which is the largest sample of students (N=1,522), Odds Ratio could only be calculated for 85% of the students and Gamma and Phi for 92% of the students each. Discrimination seemed to be most affected by zero quadrant scores leading to calculation issues: only 83% of fourth graders have valid Discrimination scores.

To examine which of these measures has the most predictive validity, the data were separated by pre and posttest. The ten measures of calibration were recalculated using only the pretest measures. Limiting the data to those students who had at least one pretest reduced the sample by three students (N=4,278). These calculable pretest measures of calibration followed the same pattern as that seen in the pre/post aggregated data (see Table 6 in Appendix A). The data were then limited to those students who had calculable values for each of the ten calibration measures, resulting in a new dataset of 3,089 students, or 72% of those with pretest data. The analysis sample was further limited to the 3,033 students who had complete demographic information (98% of the sample of 3,089 students). The resulting sample of students had data in each quadrant within the 2x2 contingency table. Student-level descriptive statistics on this

sample are presented in the right half of Table 2.5. Calibration measures from this sample are provided on the left side of Table 2.7.

{Insert Table 2.7}

As a first step, zero-order correlations were calculated to compare each of the ten measures, pretest and posttest accuracy, and pretest confidence (without regards to accuracy). These correlations are shown in Table 2.8. With one exception (Kappa and Sensitivity), all measures of calibration were correlated to levels of statistical significance of $p < .05$, with correlations ranging from .07 (Sensitivity and Phi) to a perfect correlation between G Index and Simple Match. Sensitivity had low correlations (absolute values ranging from .01 to .23) with all measures other than Specificity, with which it had a strong inverse correlation of $-.73$.

{Insert Table 2.8}

Table 2.9 displays the results from regressions of posttest accuracy on pretest accuracy and the measures of calibration, separately. Full tables with results from control variables (gender, grade, race, language and free/reduced priced lunch status, and number of questions completed) are available in the Appendix (Tables 7a and 7b). In Model 1, before the calibration measures were added, pretest accuracy and student demographics explained 69.7% of the variance in posttest accuracy. Adding an individual measure of calibration brought this, at most, to 70% of the variance as is seen in Model 7. Of the single-measure models, the Gamma model explained the most variance and also had the largest standardized regression coefficient, at 0.057. This indicates that a one standard deviation increase in Gamma was associated with less than one tenth of a standard deviation increase in aggregate posttest accuracy with pretest accuracy controlled. The combined Sensitivity/Specificity model produced a slightly larger R-squared

than the Gamma model, explaining 70.2% of the variance ($\beta=0.109$, Sensitivity, $\beta=0.074$, Specificity).

{Insert Table 2.9}

Limiting the sample to only those students who had all ten measures calculable may have biased the dataset. As an alternative analysis, the data were modified to ensure that all students with at least one valid pretest would have data in all four quadrants before the ten measures were calculated. To do this, a 1 was added to each quadrant—the right side of Table 2.7 presents the means and standard deviations from the ten measures calculated after this adjustment. Absolute differences between the sides of Table 2.7 were small (largely below .10, except for Odds Ratio), but in standard deviation units, ranged from less than 2/10ths of a standard deviation (e.g., Simple Match, G Index) to 4/10ths of a standard deviation (e.g., Odds Ratio). To determine whether these differences influenced the predictive validity of each measure, the regression of posttest score on pretest accuracy, calibration, and controls was conducted for these newly calculated measures.

Not all of the 4,278 students with pretest data also had demographic data, and so, as in the prior analyses, the data were limited to those with non-missing data on the demographic covariates, resulting in a sample of 4,144 (97% of the full pretest sample). Demographic information on the sample is provided in Appendix Table 8. Regression results were similar to those from the reduced sample and are presented in Table 2.10. A model without any calibration measures explained 67% of the variance in posttest scores. As in the prior analyses, of the single calibration measures Gamma added the most explained variance, adding an additional 0.3%. Models with Kappa and Phi also added an additional 0.3%. Unlike the limited sample models, in these regressions, G Index emerged as the strongest single predictor ($\beta=0.61$), although

differences in magnitude of standardized regression coefficients were small between many of the measures: six of the measures had betas within 0.10 of each other. Replicating the prior analysis, the combined model with Sensitivity and Specificity explained more variance than a single-measure model (67.5%). Considered in a model together, Sensitivity ($\beta=0.11$) and Specificity ($\beta=0.09$) had stronger associations with posttest score than did any other measure of calibration.

{Insert Table 2.10}

Discussion

Zero quadrants. This study set out to answer two research questions: (1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments? and (2) Among these measures, which display the greatest predictive validity? These questions were answered with data rarely used in comparisons of multiple measures of calibration: data from authentic learning tasks with a large number of questions and a large sample size. Even in the preliminary analyses, differences were apparent between these data and simulated data often created for measurement comparison studies. The students taking the quizzes within ST Math did not have accuracy and confidence judgments that were evenly distributed among cells B through D. Replicating studies meant to approximate realistic conditions (see Nietfeld et al., 2006; Schraw et al., 2013), the majority of responses were in cell A (56%). However, cell C (not confident and correct) appeared the least often (8%), indicating that few of the student responses displayed underconfident patterns. Prior research indicated concern that cell D (not confident and incorrect) would be the option most likely to remain unchosen by participants (see Schraw et al., 2011), but in my comparison of zero quadrants, cell C was the most likely cell to be left empty.

A zero in at least one quadrant affected 19% of the students using the largest possible sample of questions and participants (approximately 156 for each student). Logistic regression

results indicated that for this sample, tests of over 300 questions would be needed to avoid zero quadrants and the resulting measure distortion. This is a far cry from the 25 questions suggested by Schraw and colleagues (2011). Schraw and colleagues based this suggestion off of data simulated to replicate moderate difficulty (75% accuracy), and indicated that the more difficult the test, the more equal the distribution among the quadrants and the less likely zero quadrants would be. Based on the accuracy of the current sample, the ST Math quizzes appeared *more* difficult than Schraw's simulated data (64% accuracy), leaving a larger number of responses available for distribution in quadrants B through D. However, the majority of responses not within quadrant A (55% of the remaining responses) were in quadrant B (confident and incorrect), indicating strong overconfidence among the student participants. This overconfidence is typical in young students (Pajares & Kranzler, 1995; Pressley, Levin, Ghatala, & Ahmad, 1987). As Schraw and colleague's recommendation was based on simulated data intended to approximate adult behavior, it may not be applicable to measures of calibration in children.

It could be that the age of the children may not be the only thing causing these disparate results. The logistic regressions looked at likelihood of zero quadrants based on number of questions completed. Progress through the curriculum and completion of questions was related to a number of student characteristics (see Appendix A, Table 1). It is possible that this progress could have also been related to characteristics such as math proficiency or familiarity with the math content or format within ST Math—things that affect the students' ability to marshal metacognitive resources and make accurate confidence judgments (see Alexander & Murphy, 1999; Kruger & Dunning, 1999). Additionally, because of the structure of ST Math, the full sample included more questions from the start of the curriculum, reducing the external validity of the findings. As a step toward removing this confound and increasing external validity, sample

tests of varying question lengths were created through random selection of responses from among those students who had completed at least 200 questions. By limiting the data to only those students who had completed 200 questions, I could look within a group of students more likely to be similar to each other, and by drawing the random datasets, the difficulty of the questions was more randomly distributed. This allowed me to look at samples of small-sized quizzes (e.g., 25 questions) without having to rely on questions from a small sample of objectives that may have been easier or harder than the other objectives. Examination of these data suggested similar patterns to those observed in the data overall: quadrant C appeared the most problematic, with 40% of students missing data from this quadrant in 25-question quizzes. At 150 questions, quadrant C remained the most problematic: averaged across the grades, 15% of the students had a zero in quadrant C and 10% had a zero in quadrant D. The presence of zero quadrants decreased with the addition of more questions. Despite the suggestion from the logistic regression results indicating that zero quadrants would be eliminated at around 300 questions, I cannot say for certain that with this population and subject-matter, zero quadrants would be eliminated even at 1,000 questions, the number typically used in simulation experiments to approximate a test assumed to be problem-free.

Measure calculation. Even at test lengths of 150 questions—unreasonably high in light of typical calibration research, zeroes in quadrants were likely to be a problem. However, not all zeroes would result in undefined measures. For example, Gamma could be calculated with zeroes in quadrant C or D, as long as both were not missing.² Using all the available data, I was able to calculate all ten measures for most students. As suggested by the literature, both Gamma and Discrimination suffered from undefined values. For Gamma, between 4 and 9% of the cases,

² The ability to calculate the measure does not preclude distortion of the measure due to one zero quadrant (see Kuch, 2012).

depending on grade, were undefined or otherwise incalculable (compare with only 0.1% in a 20-item test of similar difficulty in Schraw et al., 2011). As in Fuchs (2012), there were more undefined values for Discrimination than for Gamma and the other measures. For second and fifth graders, over 20% of the values for Discrimination were undefined. This is close to the 17% Fuchs (2012) found for 20-item tests, but the majority of the test-takers in my study took well over 20 questions: fewer than 3% of the students in these data completed fewer than 20 items; the majority completed more than 150 items each.

Predictive validity. To my knowledge, this is the first study to compare the relative predictive validity of these ten commonly used measures of calibration using authentic education data. The ten measures are assumed to be correlated, except for Sensitivity and Specificity, which are assumed to be orthogonal and have proven such in simulated data (Schraw et al., 2013). Within these data, this was not the case. Sensitivity and Specificity were inversely and statistically significantly correlated and this correlation was relatively strong. Sensitivity and Specificity were more highly correlated with each other than with any of the other measures—although Specificity had moderate to strong correlations with Kappa, Phi, and Discrimination. Other researchers have advocated for Gamma as the gold standard measure of calibration, partly because of its assumed correlation with other measures (see Nelson & Narens, 1990; Schraw et al., 2013). Gamma did have strong correlations with all measures except for Sensitivity and Specificity, but it was not alone—many of the measures were as highly intercorrelated.

Given this high degree of association, their similar levels of predictive validity may not be surprising. What may be surprising is the small amount of variance in posttest accuracy explained by calibration measures. Zero-order correlations between calibration and achievement were in line with prior research (e.g., Barnett & Hixon, 1997; Desoete & Roeyers, 2006).

However, in much of this prior work, the same test is used to measure calibration and achievement (e.g., Bol et al., 2010), or correlations between pretest calibration and posttest performance are examined without controlling for pretest performance (e.g., Barnett & Hixon, 1997). If the accuracy of metacognitive judgments is indicative of a regulatory process not entirely subsumed by prior achievement, it should uniquely contribute to future mathematics performance net of prior performance. There *was* unique variance in posttest achievement explained by pretest calibration, but although statistically significant, beta values were mostly under .10. In line with the Nelson and Narens (1990) model, Gamma was the strongest singular predictor by a very small margin. However, the model combining Sensitivity and Specificity had the largest explained variance and the highest beta values. This is in agreement with Schraw and colleagues' (2013) suggestion that Sensitivity and Specificity represent unique aspects of monitoring (see also Feurman & Miller, 2008). Within these data, it appears that a model that accounts separately for students' knowledge of what they do know (Sensitivity) and what they don't know (Specificity) is more powerful than one that includes a measure that conflates the two. It is important to note that although this combined model produced the highest R-squared values in the current study, this was despite the relatively strong correlation of these two measures (cf Schraw et al., 2013).

Correcting for zero quadrants. In a replication of the regression analyses, a 1 was added to each quadrant to ensure that all measures were calculable and that zero quadrants did not otherwise distort the values of the calibration measures. I followed the procedure suggested in Hautus (1995) and added a value to each quadrant instead of only to the missing quadrants. Differences in the means of measures between this sample altered for non-missing data and the sample limited to only non-missing participants was not negligible (close to 4/10ths of a standard

deviation for Odds Ratio, Gamma, and Discrimination). These differences did not translate to large differences in predictive validity between measures, however. Betas and R-squared values between the models were close, and, as in the limited sample analysis, the model including both Sensitivity and Specificity explained the most variance and had the largest standardized regression coefficients. Prior comparisons of methods to eliminate distortion from zero quadrants relied upon tests of 1,000 questions to simulate the actual sample means of calibration measures (e.g., Nietfeld et al., 2006; Schraw et al., 2011). Within my data, I cannot say whether the limited sample with unaltered data or the full sample with altered data is closer to the *true* values of calibration. I can only note that within both, a model including Sensitivity and Specificity together explained the most variance in achievement gain, and that it appears Sensitivity and Specificity, when considered together, are better predictors of achievement gain in elementary mathematics than the other measures examined.

Limitations. The greatest strength to this study, that the data were taken from an authentic learning task completed by real students, is also a limitation. Because the data were real and suffered from real-world problems, I was unable to calculate the true values of each of the measures as is done in large simulation studies. Had a test of 1,000 questions been administered to the students in this dissertation it may have been possible to make comparisons similar to those conducted in simulated studies, but such a test is impractical in a single administration. If it were administered in smaller chunks over the course of a year it would be likely to contain the same types of missingness as was found in the data herein. Although the quizzes administered within ST Math allowed me to look at data of a type and scope not studied previously within the calibration literature, the results may be limited to similar populations and materials. There are reasons to believe there are domain and age differences in calibration (see

Bong, 1999; Jonsson & Allwood, 2003), and these differences may extend to the calculability and predictive validity of the different measures.

Conclusion

Prior recommendations regarding the measurement of calibration are based largely on simulated data. Data in this dissertation, taken from student interactions with authentic mathematics learning tasks, do not behave as simulated data do: distribution among the four quadrants is not even and patterns of missingness do not mirror those found in simulated studies. These differences have real implications for the calculability of many of the measures commonly used in calibration research. Researchers may wish to avoid measures like Gamma or Discrimination and instead rely upon measures more robust to missing quadrants, such as G Index. Outside of practical considerations, selection of a measure can also be guided by predictive validity. Results of this study supported assertions by Schraw and colleagues (2013) that Sensitivity and Specificity, when used together, should best represent metacognitive accuracy and should therefore be the most powerful predictors of achievement that relies upon SRL. These two measures have a long history of use in clinical research, but until recently were not used in the measurement of calibration within educational settings. Although the findings herein may recommend their use, more work is needed to understand the actual processes underlying determinations of confidence and uncertainty, especially in light of the high correlation between the measures within these data.

References

- Alexander, P. A., & Murphy, P. K. (1999). Nurturing the seeds of transfer: a domain-specific perspective. *International Journal of Educational Research*, 31(7), 561–576.
doi:10.1016/S0883-0355(99)00024-5
- Allwood, C. M., Jonsson, A.-C., & Granhag, P. A. (2005). The effects of source and type of feedback on child witnesses' metamemory accuracy. *Applied Cognitive Psychology*, 19(3), 331–344. doi:10.1002/acp.1071
- Barnett, J. E., & Hixon, J. E. (1997). Effects of Grade Level and Subject on Student Test Score Predictions. *The Journal of Educational Research*, 90(3), 170–174.
doi:10.2307/27542087
- 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(5), 372–382. doi:10.1016/j.learninstruc.2009.03.002
- 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(2), 133–151. doi:10.2307/20152656
- Bol, L., Riggs, R., Hacker, D. J., Dickerson, D.L., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education*, 21, 81-96.
- Bong, M. (1999). Personal Factors Affecting the Generality of Academic Self-Efficacy Judgments: Gender, Ethnicity, and Relative Expertise. *The Journal of Experimental Education*, 67(4), 315–331. doi:10.1080/00220979909598486

- California Department of Education (2011). California English Language Development Test [Test website]. Retrieved from <http://www.cde.ca.gov/ta/tg/el/>
- California Department of Education (2009). California STAR Report. Retrieved from <http://star.cde.ca.gov/star2009>
- Chen, P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77–90.
- Desoete, A., & Roeyers, H. (2006). Metacognitive Macroevaluations in Mathematical Problem Solving. *Learning and Instruction, 16*(1), 12–25.
- Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice, 14*(5), 930–933. doi:10.1111/j.1365-2753.2008.00984.x
- Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 429-455). New York: Psychology Press
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51. doi:10.3758/BF03203619
- Huff, J., & Nietfield, J. (2009). Using Strategy Instruction and Confidence Judgments to Improve Metacognitive Monitoring. *Metacognition and Learning, 4*(2), 161–176.
- Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*(4), 559–574. doi:10.1016/S0191-8869(02)00028-4

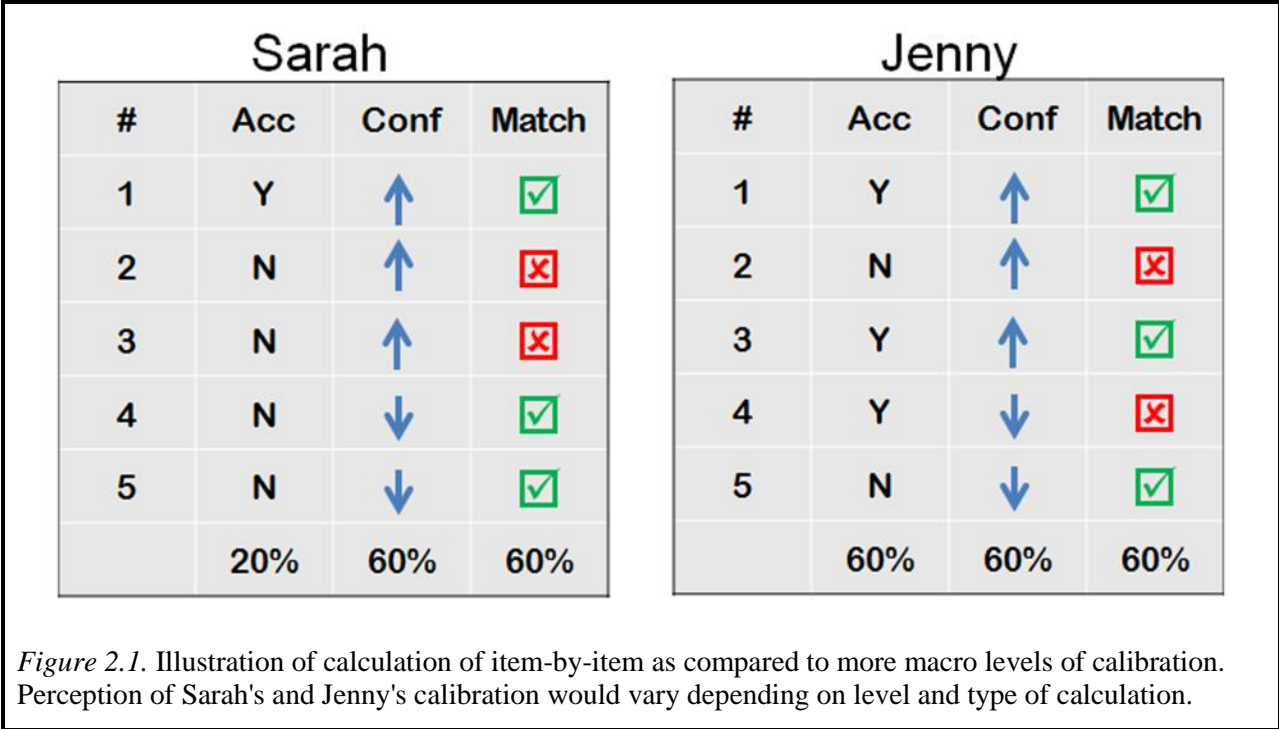
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3), 217–273. doi:10.1016/0001-6918(91)90036-Y
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Kuch, F. (2012). A comparison of bias in four measures of monitoring accuracy (Doctoral dissertation). Retrieved from <http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=2586&context=thesesdissertations>
- Lyons, K. E., & Ghetti, S. (2011). The Development of Uncertainty Monitoring in Early Childhood. *Child Development*, 82(6), 1778–1787. doi:10.1111/j.1467-8624.2011.01649.x
- Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of “nonparametric” measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, 3(2), 164–170. doi:10.3758/BF03212415
- Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 35(2), 509–527. doi:10.1037/a0014876
- Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition*, 38, 441–451. doi:10.3758/MC.38.4.441

- Miller, J. (1996). The sampling distribution of d' . *Perception & Psychophysics*, 58(1), 65–72.
doi:10.3758/BF03205476
- Nelson, T. O. & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings.
In Gordon H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 26, pp.
125–173). Academic Press. Retrieved from
<http://www.sciencedirect.com/science/article/pii/S0079742108600535>
- Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo Comparison of Measures of
Relative and Absolute Monitoring Accuracy. *Educational and Psychological
Measurement*, 66(2), 258–271. doi:10.1177/0013164404273945
- 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(3), 213–228. doi:10.1080/00220973.1997.9943455
- Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., ... Martinez, M.
E. (2014). A randomized trial of an elementary school mathematics software
intervention: Spatial-Temporal (ST) Math. *Journal of Research on Educational
Effectiveness*, online first. doi:10.1080/19345747.2013.856978
- Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem.
Applied Cognitive Psychology, 9(4), 321–332. doi:10.1002/acp.2350090405
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring.
Metacognition and Learning, 4, 33–45. doi:10.1007/s11409-008-9031-3
- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten
commonly used calibration scores. *Learning and Instruction*, 24, 48–57.
doi:10.1016/j.learninstruc.2012.08.007

Schraw, G., Kuch, F. & Roberts, R. (2011, April). Bias in the gamma coefficient: a Monte Carlo study. In *Calibrating Calibration: Conceptualization, measurement, calculation, and context*. Symposium conducted at the Annual meeting of the American Educational Research Association, New Orleans, LA.

Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95(1), 66–73.
doi:<http://dx.doi.org/10.1037/0022-0663.95.1.66>

Tobias, S. & Everson, H. T. (1998, April). *Research on the assessment of metacognitive knowledge monitoring*. Paper presented at the annual convention of the American Educational Research Association, San Diego.



A. Confident & Correct	B. Confident & Incorrect
C. Not Confident & Correct	D. Not Confident & Incorrect

Figure 3. 2x2 contingency table expressing the relations between accuracy and confidence.

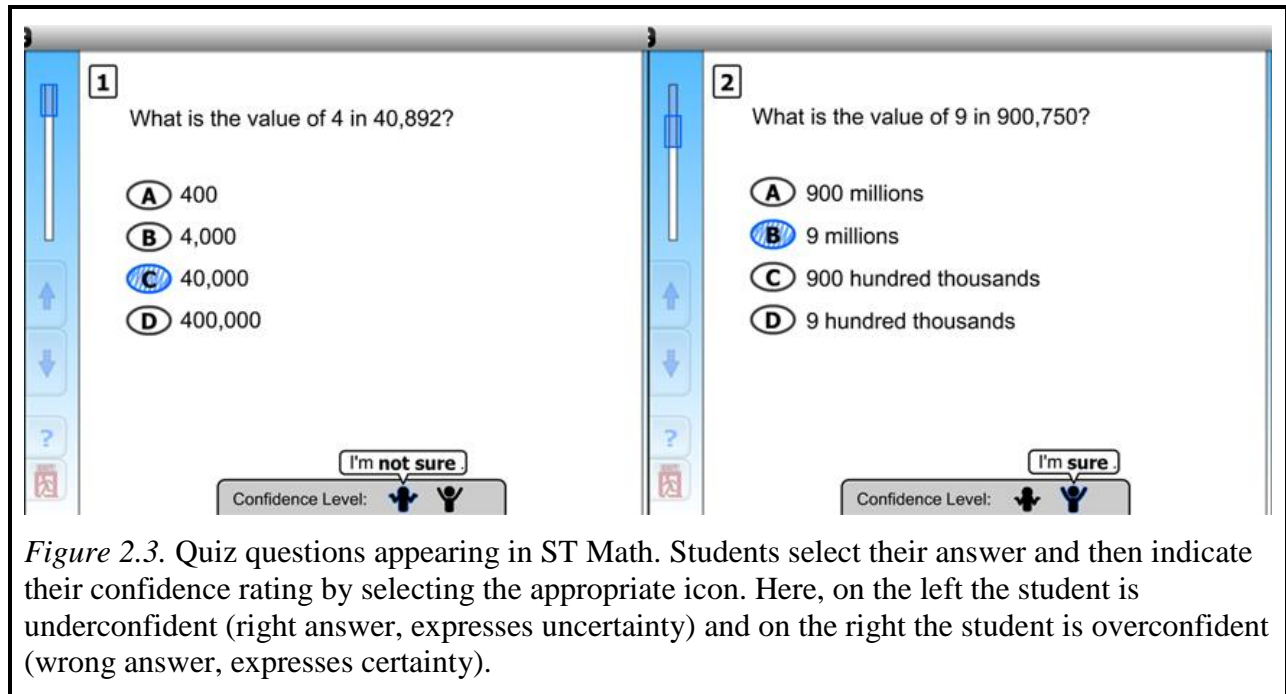


Figure 2.3. Quiz questions appearing in ST Math. Students select their answer and then indicate their confidence rating by selecting the appropriate icon. Here, on the left the student is underconfident (right answer, expresses uncertainty) and on the right the student is overconfident (wrong answer, expresses certainty).

A. Confident & Correct 56%	B. Confident & Incorrect 24%
C. Not Confident & Correct 8%	D. Not Confident & Incorrect 12%

Figure 2.4. Distribution of combinations of confidence and accuracy within the actual ST Math quiz data. Compare with Schraw (2013) simulated data where 62.5% of data was in cell A and 12.5% each in cells B through D.

Table 2.1

Common Indices of Agreement from 2x2 Contingency Tables

Index	Formula
Sensitivity	$A/(A + C)$
Specificity	$D/(B + D)$
Simple Matching	$(A + D)/(A + B + C + D)$
G Index or Hamann coefficient	$(A + D) - (B + C)/(A + B + C + D)$
Odds Ratio	AD/BC
Goodman-Kruskal Gamma	$(AD - BC)/(AD + BC)$
Kappa	$2*(AD - BC)/[(A + B)(B + D) + (A + C)(C + D)]$
Phi	$(AD - BC)/[(A + B)(B + D)(A + C)(C + D)]^{1/2}$
Sokal Reverse	$[1 - [(A + D)/(A + B + C + D)]]^{1/2}$
Discrimination (d')	$z(A/(A + C)) - z(B/(B + D))$

Note. Formulas as represented in Schraw et al., 2013.

Table 2.2
Comparison of Sample Descriptives to County and State

	Dissertation		County Mean/Percent	California Mean/Percent
	Sample Mean/Percent	Count		
Math CST	373.32	3,072	396.46	382.24
ELA CST	336.77	3,077	364.50	343.28
Male	52%	4,147	50%	49%
Free/Reduced Lunch	81%	4,147	46%	57%
Hispanic	85%	4,147	47%	50%
White	8%	4,147	31%	26%
Asian	3%	4,147	14%	9%
Other Race	3%	4,147	8%	23%
Eng Language Learner	65%	4,146	39%	32%
N	4,281		110,402	1,401,811

Note. Column 1 is calculated from available data within the sample. Demographic data were only present for the specified number of students. County and California data aggregated for grades two through four in 2008-2009 from the California STAR reporting website: <http://star.cde.ca.gov/star2009>.

Table 2.3

Mean Number of Questions Answered & in Each of the Calibration Categories, Aggregated Quiz Questions

	2nd (N=915)			3rd (N=812)			4th (N=1,522)			5th (N=1,032)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
A. Confident & Correct	90.85	53.88	0-235	104.30	62.04	0-245	80.85	52.26	0-235	79.63	50.03	0-230
B. Confident & Incorrect	39.87	30.34	0-165	42.87	28.30	0-182	38.65	26.40	0-147	32.96	23.60	0-131
C. Unconfident & Correct	14.02	20.50	0-167	12.87	19.75	0-201	12.92	15.03	0-136	10.03	13.17	0-95
D. Unconfident & Incorrect	17.32	19.61	0-126	17.84	19.60	0-132	22.46	23.37	0-167	14.76	16.39	0-118
Total Questions Answered	162.07	76.30	7-270	177.88	76.49	14-280	154.88	71.25	5-270	137.38	66.13	5-230

Table 2.4

Odds Ratios and Marginal Effects from Logistic Regression of Zeroes in Quadrants C and D by Total Number of Questions

N=4,281	C. Not Conf. & Correct		D. Not Conf. & Incorrect	
	Odds Ratio	Marg. Effects	Odds Ratio	Marg. Effects
Total No. Questions	.994*** (.001)	-.001*** (.0001)	.993*** (.001)	-.001*** (.0001)
Grade 2	1.29* (.149)	.033* (.016)	1.395* (.203)	.028* (.013)
Grade 3	1.015 (.131)	.002 (.016)	1.002 (.169)	.0001 (.013)
Grade 5	1.055 (.119)	.007 (.014)	1.470** (.199)	.032** (.012)
Constant	.430*** (.047)		.227*** (.031)	

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

Standard errors in parentheses. Grade 4 is the reference group.

Table 2.5
Demographic Information and Descriptive Statistics for Study Subsamples

	Sample of Students Who Answered > 200 Questions					Sample of Students for Predictive Validity Analysis				
	Mean/Percent	SD	Min	Max	Count	Mean/Percent	SD	Min	Max	Count
Grade 2	26%				1,341	20%				3033
Grade 3	24%				1,341	19%				3033
Grade 4	34%				1,341	37%				3033
Grade 5	17%				1,341	23%				3033
Male	57%				1,335	49%				3033
Asian	6%				1,335	3%				3033
Hispanic	77%				1,335	85%				3033
White	12%				1,335	9%				3033
Other Race	4%				1,335	3%				3033
English Lang Learner	57%				1,334	65%				3033
Free/Reduced Lunch	70%				1,335	80%				3033
ELA CST	420.15	75.63	219	600	953	339.72	60.20	179	600	2274
Math CST	365.13	58.06	210	600	956	377.81	79.16	181	600	2269
Pretest Quiz Accuracy	0.62	0.14	0.27	1	1,341	0.57	0.14	0.2	0.98	3033
Posttest Quiz Accuracy	0.73	0.13	0.29	1	1,341	0.67	0.15	0.11	1	3033
Total Pretest Questions	121.21	11.12	93	140	1,341	83.60	33.78	5	140	3033
Total Posttest Questions	122.31	10.92	100	140	1,341	84.57	33.90	4	140	3033
N	1,341					3,033				

Note. Left section consists of those students who answered at least 200 questions across pre and posttests. The right section consists of those students who have valid data on all measures of calibration for the pretest and valid posttest accuracy data (excluding those with incalculable measures due to zero quadrants). Count column represents number of students with valid data for each variable.

Table 2.6

Ten Common Measures of Calibration Calculated for Entire Sample from All Available Data

N=4,281	2nd Grade			3rd Grade			4th Grade			5th Grade		
	Mean	SD	% Valid	Mean	SD	% Valid	Mean	SD	% Valid	Mean	SD	% Valid
Sensitivity	0.86	0.17	99.89%	0.87	0.16	100.00%	0.85	0.16	100.00%	0.87	0.15	100.00%
Specificity	0.31	0.27	99.34%	0.29	0.24	100.00%	0.36	0.27	99.67%	0.31	0.24	98.64%
Simple Match	0.66	0.12	100.00%	0.67	0.14	100.00%	0.65	0.13	100.00%	0.68	0.12	100.00%
Gamma	0.54	0.41	92.57%	0.54	0.41	96.06%	0.57	0.37	94.22%	0.53	0.42	91.38%
G Index	0.33	0.25	100.00%	0.33	0.27	100.00%	0.3	0.25	100.00%	0.35	0.25	100.00%
Odds Ratio	5.42	5.52	82.30%	5.84	6.62	86.95%	5.72	5.43	85.48%	5.32	6.07	82.46%
Kappa	0.17	0.16	99.45%	0.18	0.16	100.00%	0.21	0.17	99.67%	0.18	0.16	99.42%
Phi	0.22	0.16	92.57%	0.22	0.16	96.06%	0.25	0.16	94.22%	0.23	0.16	91.38%
Sokal Reverse	0.57	0.11	100.00%	0.57	0.12	100.00%	0.58	0.11	100.00%	0.56	0.12	100.00%
Discrimination	0.76	0.47	79.34%	0.76	0.48	83.67%	0.82	0.46	83.25%	0.77	0.43	78.59%
N	915			812			1,522			1,032		

Note. Includes all students in the sample combining questions in both pre and posttests. Ten calibration measures are calculated as in Schraw et al. (2013) based on four quadrants of agreement between accuracy and confidence. % Valid represents the percent of students for whom the given measures is calculable.

Table 2.7

Means of Ten Measures of Calibration Eliminating (Left) and Accommodating (Right) Missing Data

	Limited Sample (N=3,033)				Complete Sample (N=4,278)			
	2nd	3rd	4th	5th	2nd	3rd	4th	5th
Sensitivity	0.80 (0.17)	0.82 (0.18)	0.78 (0.17)	0.81 (0.16)	0.83 (0.17)	0.84 (0.17)	0.81 (0.17)	0.83 (0.16)
Specificity	0.41 (0.26)	0.37 (0.23)	0.44 (0.24)	0.40 (0.23)	0.33 (0.26)	0.31 (0.23)	0.36 (0.26)	0.34 (0.23)
Simple Match	0.64 (0.12)	0.65 (0.13)	0.64 (0.11)	0.64 (0.11)	0.62 (0.12)	0.63 (0.13)	0.61 (0.12)	0.63 (0.12)
Gamma	0.49 (0.33)	0.50 (0.32)	0.51 (0.30)	0.50 (0.28)	0.41 (0.33)	0.42 (0.35)	0.41 (0.34)	0.43 (0.32)
G Index	0.28 (0.24)	0.30 (0.25)	0.28 (0.22)	0.28 (0.21)	0.24 (0.24)	0.25 (0.27)	0.22 (0.24)	0.26 (0.24)
Odds Ratio	5.49 (6.78)	5.79 (7.69)	5.28 (5.81)	5.46 (11.51)	3.96 (3.90)	4.39 (4.80)	3.96 (3.84)	4.54 (8.22)
Kappa	0.21 (0.17)	0.20 (0.16)	0.23 (0.16)	0.21 (0.16)	0.16 (0.16)	0.16 (0.16)	0.18 (0.16)	0.17 (0.15)
Phi	0.23 (0.18)	0.23 (0.17)	0.25 (0.17)	0.24 (0.16)	0.18 (0.17)	0.18 (0.16)	0.19 (0.17)	0.19 (0.16)
Sokal Reverse	0.59 (0.10)	0.58 (0.11)	0.59 (0.09)	0.59 (0.09)	0.61 (0.10)	0.60 (0.11)	0.62 (0.10)	0.60 (0.10)
Discrimination	0.71 (0.54)	0.72 (0.52)	0.74 (0.50)	0.72 (0.48)	0.56 (0.49)	0.58 (0.51)	0.57 (0.50)	0.59 (0.49)
N	601	570	1118	712	915	812	1,521	1,030

Note. Standard deviations in parentheses. Ten calibration measures are calculated as in Schraw et al. (2013) based on four quadrants of agreement between accuracy and confidence. Limited Sample excludes those missing on any measure due to zero quadrants. Complete sample is based on calculations after adding 1 to each quadrant.

Table 2.8

Correlations among Measures of Calibration, Pre and Posttest Accuracy and Pretest Confidence in Reduced Sample (N=3,033)

Measure	Pretest		Posttest		Conf		Sens		Spec		Match		Gamma		G Index		OddsRa		Kappa		Phi		SokalR		
Posttest Acc.	0.81		0.63		0.53		0.29		-0.33		0.33		0.77		1.00		0.50		0.57		0.99		-0.76		-0.79
Confidence	0.63	0.63	0.29	0.29	-0.33	-0.73	0.33	0.22	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Sensitivity	0.29	0.29	-0.05*	-0.05*	-0.33	-0.73	0.33	0.22	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Specificity	-0.05*	-0.05*	0.67	0.57	0.47	0.22	0.33	0.22	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Simple Match	0.67	0.57	0.47	0.22	0.33	0.22	0.33	0.22	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Gamma	0.35	0.35	0.29	0.21	0.41	0.21	0.41	0.21	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
G Index	0.67	0.57	0.47	0.22	0.33	0.22	0.33	0.22	0.41	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Odds Ratio	0.26	0.23	0.18	0.16	0.26	0.16	0.26	0.16	0.26	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Kappa	0.27	0.26	0.11	0.01 ^a	0.66	0.01 ^a	0.66	0.01 ^a	0.66	0.76	0.84	0.76	0.84	0.76	0.84	0.76	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Phi	0.27	0.28	0.15	0.07	0.61	0.07	0.61	0.07	0.61	0.77	0.92	0.77	0.92	0.77	0.92	0.77	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79
Sokal Reverse	-0.68	-0.58	-0.48	-0.23	-0.31	-0.23	-0.31	-0.23	-0.31	-0.99	-0.75	-0.99	-0.75	-0.99	-0.75	-0.99	-0.55	-0.55	-0.57	-0.57	-0.99	-0.99	-0.76	-0.76	-0.79
Discrimination	0.34	0.33	0.23	0.16	0.51	0.16	0.51	0.16	0.51	0.79	0.95	0.79	0.95	0.79	0.95	0.79	0.50	0.50	0.57	0.57	0.99	0.99	-0.76	-0.76	-0.79

Note. All correlations are statistically significant at the $p < .001$ level except for those specified: * $p < .05$. ^a $p > .05$.

Pretest and posttest accuracy and confidence are a proportion of accurate or confident out of total test items.

Table 2.9a

Regression of Posttest Accuracy (Percentage of Items Correct) on Pretest Calibration and Accuracy for Ten Measures of Calibration: Diagnostic Efficiency & Agreement Measures

N=3,033		(1)	(2)	(3)	(4)	(5)	(6)	
		Acc. Only	Sensitivity	Specificity	Sensitivity	Specificity	Simple Match	G Index
Measure(s)	B		0.046***	-0.003	0.098***	0.047***	0.074***	0.037***
	SE		(0.009)	(0.006)	(0.014)	(0.010)	(0.018)	(0.009)
Pretest Acc.	Beta		0.052***	-0.004	0.109***	0.074***	0.056***	0.056***
	B	0.818***	0.803***	0.818***	0.789***		0.779***	0.779***
	SE	(0.012)	(0.012)	(0.012)	(0.012)		(0.015)	(0.015)
Constant	Beta	0.758***	0.744***	0.758***	0.731***		0.721***	0.721***
	B	0.147***	0.121***	0.148***	0.066***		0.121***	0.158***
	SE	(0.009)	(0.011)	(0.010)	(0.015)		(0.011)	(0.009)
R2		0.697	0.699	0.697	0.702		0.698	0.698

Table 2.9b

Association, Binary Distance, and Discrimination

N=3,033		(7)	(8)	(9)	(10)	(11)	(12)
		Gamma	Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
Measure(s)	B	0.028***	0.0004*	0.046***	0.049***	-0.081***	0.017***
	SE	(0.005)	(0.0002)	(0.010)	(0.010)	(0.022)	(0.003)
Pretest Acc.	Beta	0.057***	0.021*	0.049***	0.054***	-0.052***	0.055***
	B	0.798***	0.812***	0.804***	0.803***	0.781***	0.799***
	SE	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)	(0.012)
Constant	Beta	0.739***	0.752***	0.745***	0.744***	0.723***	0.740***
	B	0.144***	0.147***	0.143***	0.142***	0.215***	0.145***
	SE	(0.009)	(0.009)	(0.009)	(0.009)	(0.020)	(0.009)
R2		0.700	0.697	0.699	0.699	0.698	0.699

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included number of pre and posttest questions completed, grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch. Sample limited to those students who have non-missing values for each of the ten measures of calibration as described in Schraw et al. (2013).

Table 2.10a

Replication Sample: Regression of Posttest Accuracy (Percentage of Items Correct) on Pretest Calibration and Accuracy for Ten Measures of Calibration: Diagnostic Efficiency & Agreement Measures

N=4,144		(1)	(2)	(3)	(4)	(5)	(6)	
		Acc. Only	Sensitivity	Specificity	Sensitivity	Specificity	Simple Match	G Index
Measure(s)	B		0.035***	0.003	0.111***	0.061***	0.080***	0.040***
	SE		(0.009)	(0.006)	(0.015)	(0.010)	(0.018)	(0.009)
	Beta		0.036***	0.005	0.114***	0.094***	0.061***	0.061***
Pretest Acc.	B	0.871***	0.860***	0.871***	0.838***		0.819***	0.819***
	SE	(0.011)	(0.011)	(0.011)	(0.012)		(0.016)	(0.016)
	Beta	0.750***	0.741***	0.750***	0.721***		0.705***	0.705***
Constant	B	0.130***	0.108***	0.129***	0.037*		0.109***	0.149***
	SE	(0.008)	(0.010)	(0.009)	(0.015)		(0.010)	(0.009)
R2		0.670	0.672	0.670	0.675		0.672	0.672

Table 2.10b

Association, Binary Distance, and Discrimination

N=4,144		(7)	(8)	(9)	(10)	(11)	(12)
		Gamma	Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
Measure(s)	B	0.029***	0.0004	0.052***	0.057***	-0.075***	0.018***
	SE	(0.005)	(0.0003)	(0.010)	(0.009)	(0.022)	(0.003)
	Beta	0.060***	0.012	0.051***	0.058***	-0.047***	0.056***
Pretest Acc.	B	0.841***	0.866***	0.853***	0.849***	0.830***	0.844***
	SE	(0.012)	(0.012)	(0.012)	(0.012)	(0.016)	(0.012)
	Beta	0.724***	0.746***	0.735***	0.731***	0.714***	0.727***
Constant	B	0.135***	0.131***	0.130***	0.131***	0.198***	0.135***
	SE	(0.008)	(0.009)	(0.008)	(0.008)	(0.022)	(0.008)
R2		0.673	0.670	0.673	0.673	0.671	0.672

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included number of pre and posttest questions completed, grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch. Regressions on full sample of students who have at least one valid pre and posttest.

CHAPTER 3

Study 2: Within and Between Person Associations of Calibration and Achievement

A wide range of research has investigated the associations between performance and the calibration of accuracy and confidence. Mostly undertaken in the fields of education or psychology, calibration research has dealt with topics including eyewitness testimony (e.g., Howie & Roebbers, 2007), text comprehension (e.g., Maki & Berry, 1984), and class performance (e.g., Bol & Hacker, 2001). A consistent finding is that higher performers display better metacognitive monitoring, operationalized as some form of calibration (Stone, 2000; e.g., Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010; Ots, 2012; Soku & Quereshi, 2004). Conversely, poor performers are often "doubly cursed" in that not only do they perform poorly, but they are often unaware of their own poor performance, making it unlikely that they will take corrective action (Dunning, Johnson, Ehrlinger, & Kruger, 2003). This relation between monitoring and performance is theorized to operate through a system of self-regulated learning (SRL), where monitoring can alert learners to engage in control processes and allocate attention and resources where needed (Pintrich, 2004; Winne, 1995, 2004; Zimmerman, 2008). Although the calibration/performance relation is well-documented, it is often studied with a dispositional view of monitoring—examining performance differences between students who are good monitors and those who are poor monitors (e.g., Barnett & Hixon, 1997; Chen, 2002; Soku & Quereshi, 2004). This view of monitoring is not in line with the Social Cognitive view of SRL as one that varies based on interactions between person, behavior, and environment (Bandura 1986; Zimmerman 1989). Nor does it distinguish monitoring from other individual characteristics that are related to both monitoring and performance. The current study takes a novel approach to examine the relation between monitoring and performance within the same person across

multiple tasks—analyzing whether differences in calibration across tasks are associated with differences in performance. In this way the dynamic nature of calibration can be better represented and the unique contribution of calibration within the SRL system can be better understood.

Models of Regulation: Monitoring and Control

The model of regulation adopted within this study is situated within the larger frame of SRL as described by Zimmerman (1986; 1989; see Dissertation Introduction for a description). However, much of the work on calibration and its role within SRL has been situated in a line of research stemming from Flavell's (1979) and later Nelson and Narens' (1990) conceptualizations of metacognition. In particular, I draw largely on the work of Efklides and her Metacognitive and Affective Model of Self-Regulated Learning (MASRL model, 2011), as well as her other work explicating the relations between monitoring and control (Efklides, 2008; Efklides & Vlachopoulos, 2012). Although typically distinct lines of research, the Flavell and Zimmerman conceptualizations of SRL and metacognition are complementary, jointly emphasizing the role of the learner as an agent of their own learning and acknowledging the contributing role of both learner background and task characteristics.

Within the MASRL model, metacognition, motivation, and affect interact across two levels, the Person level and the Task x Person level (Efklides, 2011). The person level includes personal characteristics such as self-beliefs, ability, and person-level metacognitive knowledge (MK) and metacognitive skills (MS)—knowledge and skills that apply to a variety of tasks and a sense of when and how to apply them. The Task x Person level is where online metacognition takes place: based on an individual's experience of the task, she represents the task in a way that allows her to draw on the Person level (e.g., MK, MS, motivation) and engage control processes

in light of her metacognitive experiences (ME). These metacognitive experiences are cues from the individual's interaction with the task, such as an awareness of ease or difficulty of processing, and feelings of familiarity or confidence (Efklides, 2008).

According to Efklides (2008), as individuals engage in a task, they engage in a non-conscious implicit form of monitoring and control that responds to the demands of the task. An example of this non-conscious monitoring is seen as individuals slow down their rate of reading of difficult material even though they may not be consciously aware that the difficulty has increased. If an error is detected that cannot be resolved through this implicit regulation, a conscious level of regulation is activated. This error detection and activation of conscious regulation can be because of the difficulty of the task or because of external feedback that draws the learner's attention to the monitoring and control processes. Once monitoring information is active within working memory, the learner can take steps to exercise control: increasing resources brought to bear on the task or allocating resources differently across different elements of the task. Learners who activate conscious monitoring and who are accurate in their judgments of learning can appropriately allocate resources in the control process (see Efklides, 2008; Koriat, 2012).

In studies of metacognitive monitoring using calibration judgments, participants are likely engaging in conscious monitoring—the self-report questions ubiquitous in such studies (e.g., Boekaerts & Rozendaal, 2010; Bol & Hacker, 2001; Nietfeld, Cao, & Osborne, 2006) bring to the learner's awareness the presence or absence of discrepancies between their current and goal states of performance (see De Bruin & van Gog, 2012). With this entry into conscious monitoring, the success of the learner at regulation then in part depends on the accuracy of these judgments—those who accurately identify a discrepancy may trigger control processes (Efklides,

2008; De Bruin & van Gog, 2012). Figure 3.1 illustrates this theorized process with a flow chart derived from the research and theories of De Bruin and van Gog (2012), Efklides (2008; 2012), Koriat (2012), and Nelson and Narens (1990). The flow chart should be read from left to right and is simplified to illustrate the processes as recursive.³ After activation of conscious regulation, the individual draws on relevant personal characteristics based on her experience of the task (ME) and her characterization of the task. In the model, these personal characteristics are the applicable MK and MS as well as ability, self-beliefs, etc.—the trapezoid shape narrowing at the base represents that some of these features of the Person level may be more or less relevant to the task. For example, for a fraction task, self-efficacy for fractions would be more relevant, general mathematics self-efficacy less specifically relevant, and self-esteem still less relevant. Once monitoring triggers conscious regulation through the ME, the interaction of the Person and Task x Person levels results in a decision to engage in control, disengage from the task, or return to non-conscious monitoring. Once the decision is made to engage in control, the choice of control activity is informed by these same Person and Person x Task levels, both filtered through the learner's attributions for the particular discrepancy of which they have become aware. For example, while taking a test, a learner who is not confident that she has the correct answer but who feels she can arrive at the correct answer will engage in control processes. If she attributes her incorrect answer to not properly understanding the question, it is through this attribution that she will bring to bear Person level characteristics, such as her knowledge of strategies for clarifying instructions or her self-efficacy for solving problems in general or problems of this type. Through this process she chooses a control activity: in this case she may reread the question using strategies she has identified as relevant and potentially helpful.

³ Consistent with SRL theory (Zimmerman, 1989) and previous models illustrated by Efklides (2008, 2011), the processes within the Figure 3.1 model are theorized to be non-recursive: control activities and attributions feed back into Person and Task x Person characteristics. For parsimony, this is omitted from the model shown.

{Insert Figure 3.1 }

As typically studied, calibration accuracy (as one aspect of ME) is measured with a task, such as a postdiction of the percentage of questions correct on an Educational Psychology exam (e.g., Bol, Hacker, O'Shea, & Allen, 2005). This measure of calibration is then related to performance on the same exam (e.g., Bol et al., 2010), a task from an unrelated or tangentially-related domain (e.g., Ots, 2012), or a test in the same domain much later in time (e.g., Rinne & Mazocco, 2014). This makes it difficult to understand the function of calibration within the ecology of SRL: the link between student calibration on knowledge recall questions and GPA may indicate a disposition toward metacognitive monitoring, but it does not indicate whether this monitoring may serve to enhance regulation of learning from task to task or within the same task.

Additionally, such a study confounds calibration with other aspects of conscious regulation. When the only aspect of conscious regulation measured is calibration, what may be misconstrued as a link between calibration and performance may in actuality be a link between any number of personal characteristics and performance. There is both theoretical and empirical evidence that measures of calibration may largely reflect stable personal characteristics that have little to do with metacognitive monitoring (Pieschl, 2009; Scheck, Meeter, & Nelson, 2004; Zhao & Linderholm, 2008). Zhao and Linderholm (2008) present a theory wherein individuals, in making monitoring judgments, first anchor their judgment with expectations based on past experiences from potentially unrelated tasks and then adjust based on features of the actual task, ending with a judgment that, despite adjustment, is biased toward stable personal characteristics without adequately addressing task-specific considerations. Other researchers have relied upon the stability of calibration across tasks as indicative of a stable monitoring characteristic (e.g.,

Mengelkamp & Bannert, 2010). Without disentangling indicators of monitoring from other personal characteristics, the true value of accurate calibration is unlikely to be revealed.

The Current Study

Relying on a model of conscious regulation as described by the flow chart in Figure 3.1, the current study seeks to determine a more true association between calibration and achievement. The context for the study is an online mathematics learning environment, Spatial Temporal Mathematics (ST Math), created by the MIND Research Institute. Students participating in ST Math proceeded through a grade-level-specific curriculum, divided into 21 to 25 objectives, depending on grade level. Objectives covered included mathematics topic areas such as “Multi-Digit Multiplication” and "Linear Functions and Equations." Appendix B, Table 1 provides a description of each of the objectives within the ST Math curriculum, divided by grade. The content was ordered to approximate the progression of content within a typical mathematics class, but was not aligned with pacing guides or other curricular materials.

Each objective within ST Math was prefaced with a 5 to 10 question pretest on objective-relevant content (examples provided in Appendix B). Within the pretest, once students selected their answer for a question, they were prompted to indicate their confidence in this answer by selecting a cheering icon to represent certainty or a shrugging icon to represent uncertainty (see Study 1 Figure 2.1). Students were then allowed to review their answers and confidence ratings on each question before beginning the main objective content. The main content consisted of a number of learning games leveled by difficulty. Within each game, students solved puzzles to help the ST Math penguin, Jiji, proceed from left to right across the screen through the use of mathematical problem solving to remove impediments in Jiji's path. Students had to correctly complete 80% of the puzzles within each level to move on to the next level; however, students

were able to replay levels as desired and otherwise proceeded at their own pace: there was no time-limit within the game for the completion of a given puzzle, level, or objective. At the end of each objective, students took a posttest quiz mirroring the content of the pretest quiz. The posttest quiz problems were structured in the same way as the pretest quiz problems: students were asked for confidence judgments on each problem and were allowed a post-quiz period of review.

As students took the pretest quizzes and made judgments of confidence, their attention was directed toward monitoring, presumably passing the threshold into conscious regulation as illustrated in Figure 3.1. At this point, students had access to monitoring information as they proceeded to the main learning phase, where, if they were accurately calibrated and in need of performance improvements, they would engage control processes to regulate their learning (such as controlling their attention or replaying games). Because the content of each objective (and objective quiz) varied, confidence judgments and their accuracy were also likely to vary, due in part to features of the task and interactions between person and task. If students were better able to engage in control processes during the gameplay in objectives in which they were more accurately calibrated, and these control processes successfully influenced learning, then pre to posttest gains on these objectives would be larger than pre to posttest gains on objectives in which they were more poorly calibrated. By looking only at the joint variation in calibration and performance *within* each student, the Person level of conscious regulation as represented in Figure 3.1 can be eliminated from the model, reducing bias on our estimate of calibration's association with performance. The non-monitoring-aspects of the Task x Person level still remain and are potential confounds to our estimate as are any task-specific personal characteristics (e.g., self-efficacy for ST Math problems); however, because the content across

objectives shares a large number of features (in that it is grade-level mathematics problem-solving presented and tested within the same format), it can be assumed that some aspects of the Task x Person level are also controlled. Arguably one of the most important features of the Task x Person level, student familiarity with the particular task (see de Bruin & van Gog, 2012; Dunning et al., 2003) can be controlled by adding a pretest covariate in the model.

Measures of calibration. Isolating calibration from other aspects of conscious regulation is one step toward better understanding the dynamic nature of monitoring within SRL. Certain operationalizations of monitoring may also shed light on different processes and the different ways in which monitoring can lead to control and improvements in performance. Assuming a general monitoring ability that can be measured as a single construct, learner use of this ability to identify which content they know and which content they don't know can enable the most efficient application of resources (Nelson & Narens, 1990; Schraw, Kuch, & Gutierrez, 2013). Learners could direct attentional resources away from material already mastered and toward material that has yet to be mastered. In experimental studies, cognitive scientists have demonstrated that individuals do indeed allocate more study time to items they deem as more difficult to learn (e.g., Nelson & Narens, 1990), although some research has found that under certain circumstances individuals will choose easier items first (Theide & Dunlosky, 1999). Assuming such a singular process, measures such as Gamma or Discrimination may best capture the process of forming metacognitive judgments (see Study 1 for a description of these measures). However, if instead of a single monitoring ability, learners use distinct processes to make judgments of confidence and judgments of uncertainty, more than one measure would be necessary (Schraw et al., 2013).

There is theoretical support for a two-process model. It is a consistent finding that poor performers display overconfidence; however, it is also a consistent finding that although better calibration is associated with better performance, the best performers tend to display underconfidence (Stone, 2000). The top-down process by which individuals make confidence judgments (see Zhao & Linderholm, 2008) may illuminate this finding. As individuals draw on general self-beliefs or prior experiences to make their judgments for a given task, those who are unfamiliar with the domain or topic of study will not have access to the information necessary for them to adjust their judgments in consideration of the task, and will therefore likely underestimate the task demands, leading to overconfidence (Kruger & Dunning, 1999). The converse may be true: those with more prior knowledge may have an abundance of resources upon which to draw and may overthink the problem, causing underconfidence. Additionally, metacognitive experiences at the Task x Person level feed back into more stable beliefs at the Person level (Efklides, 2008), and so it may be protective, especially for those who feel threatened, to bolster their more general sense of self-worth with high confidence judgments (see Ots, 2012). Ots (2012) also offers evidence that high performers may underestimate as a form of defensive pessimism.

Schraw and colleague's' (2013) analysis of ten measures of calibration using simulated data supported a two-process model, finding that including the measures of Sensitivity (proportion confident when correct) and Specificity (proportion uncertain when incorrect) in a model together best accounted for variance within the data. Findings within Study 1 of this dissertation also support the use of these measures: although differences between models were small, a model including Sensitivity and Specificity together explained the most variance within the data, and these two measures jointly had the greatest predictive validity for performance.

Developmental factors in monitoring and control. In considering Person level characteristics, which may be related to both monitoring and achievement, it may be beneficial to examine the moderating effects of those characteristics that are directly observable. Age is one such characteristic hypothesized to influence achievement and to influence both monitoring and control; study of the age-related development of metacognition has long been a topic of interest (see Flavell, 1979). Prior research has noted improvement in metacognition and regulation as children age, especially across elementary school, and has provided evidence that monitoring processes may be responsible for these changes (e.g., Howie & Roebbers, 2007; Pressley & Ghatala, 1990; Pressley, Levin, Ghatala, & Ahmad, 1987). Young children may not attend to important features of tasks (Markman, 1977) and may tend toward wishful thinking or overconfidence (Desoete & Royers, 2006; Saxe & Sicilian, 1981; Schneider, 2002). There is evidence however that even very young children can make accurate confidence judgments, especially in situations where the task is simple and/or well-known and the directions are clear (de Bruin & van Gog, 2012; Ghetti, Hembacher, & Coughlin, 2013; Roebbers, 2002; Schneider, 2002).

There may be age-related differences in both the ways children make monitoring judgments and in how they use them toward control processes. Children and adults both make more accurate judgments after a delay (Schneider, Visé, Lockl, & Nelson, 2000); however, they may respond differently to feedback (Newman & Wick, 1987). Children's ability to monitor metacognition may especially suffer from working memory demands in complicated tasks (Ghetti et al., 2013; Hacker, Dunlosky, & Graesser, 1998). Even with the same mean-level of calibration accuracy, younger children may not be as able to use information from metacognitive monitoring to influence control processes (Destan, Hembacher, Ghetti, & Roebbers, 2014). As

with accuracy of metacognitive judgments, this translation of monitoring to control may depend on familiarity with and understanding of the task (de Bruin & van Gog, 2012).

Research questions. I have presented a model of conscious regulation wherein accuracy of metacognitive monitoring, as calibration, affects performance on a task through the engagement of control processes. Within this model, calibration can be disentangled from other features of conscious regulation by examining associations between calibration and performance within the same student across related tasks. This study explores these associations within an online mathematics learning environment, ST Math, and asks (1) Do students (within ST Math) make greater pre to posttest gains when better calibrated at pretest? If there are statistically significant associations between calibration and performance within student, then there is evidence that elementary students are using ME to enact control processes and influence their learning. If there remain associations between calibration and performance between students, then there is evidence that there is a stable metacognitive monitoring trait or that another stable Person level characteristic is associated with calibration and performance.

To answer this question, calibration is operationalized as Sensitivity, proportion confident when correct, and Specificity, proportion uncertain when incorrect. Representing these aspects of monitoring separately is in keeping with Schraw et al. (2013) and with the findings of Study 1 of this dissertation, and will also allow an examination of the potentially different processes surrounding accurate judgments of confidence and uncertainty. Within student differences in associations between these two measures and performance may indicate that they exert differential influences on control and thus performance, or may indicate that they are biased by different elements at the Task x Person level. For example, if influence on control differs between the measures, confidence of correct answers could allow students to operate more

efficiently and proceed more quickly through the content, whereas uncertainty in incorrect answers may indicate to students where they need to direct their attention or replay levels. Differences in associations *between* students may indicate that stable Person level characteristics associated with both calibration and performance differ between the two measures.

As a second research question, I ask (2) Does calibration and the benefit from calibration vary depending on student grade-level? A grade-related improvement in the mean-level of calibration may indicate that the older children display better metacognitive monitoring in mathematics problem solving. It may be that mean-levels of calibration are equal across the grades but older students *use* metacognitive judgments differently through control processes. If this is the case, the associations between calibration and performance will vary by grade.

Method

Sample. The sample for this study consisted of approximately half of the second through fifth graders at 18 schools in Southern California. The participating schools were largely Hispanic (85%) and low-income (80% eligible for free/reduced lunch), and on average, lower performing than the other county and state schools (see tables in Study 1). The schools had been randomly assigned to receive Spatial Temporal (ST) Math, an online spatially-based mathematics curriculum, in either second and third or fourth and fifth grades. Because the year of data collection for this paper (2010) is the second year of the study, there was no treatment/control variation in fourth grade. Therefore, all fourth graders in the participating schools were included in this sample. There were 4,137 students in grades 2 through 5 who used ST Math within the study schools. The current analyses were limited to the 3,912 students (95%) who were using their on-grade curriculum (excluding fifth graders using fourth grade curriculum, for example), and who had completed at least two of the 20+ objectives within a

year's ST Math curriculum. Comparisons between the samples and descriptive statistics on both can be seen in Table 3.1. Excluded students are more likely to be in fifth grade and male; no other differences between excluded and retained students rose to the level of statistical significance.

{ Insert Table 3.1 }

Procedure. Students played ST Math for 45 minutes at a time during twice weekly visits to the computer lab throughout the academic year. Student selections on multiple choice quiz questions along with their ratings of confidence were collected and compiled by MIND and provided to the author, who was able to match them with state identifiers and demographic information.

Measures.

Quiz data. For each objective, accuracy is represented separately for pre and posttest quizzes as percentage correct. Calibration is operationalized as recommended by Schraw et al. (2013), with Sensitivity (percent of correct items where students noted confidence) and Specificity (percent of incorrect items where students noted uncertainty), based on the distribution of data within the 2x2 contingency table of confidence and accuracy (See Study 1 Figure 2.2). To accommodate students who did not have data in each combination of confidence and accuracy, .01 was added to each quadrant before Sensitivity and Specificity were calculated for each quiz. Because of this addition, the range of Sensitivity and Specificity were changed from the standard zero to one range to a range just over zero to approaching one. A score of .5 can be considered *neutral* in this case, but can be arrived at in two ways: by not having any correct or incorrect answers, or by not indicating appropriate confidence on half of the correct/incorrect answers. For example, for Sensitivity, a student who did not get any items

correct would have $.01/(.01+.01)$ or $.5$, as would a student who indicated confidence on two items but had actually answered four items correctly: $2.01/(2.01 + 2.01)$ or $.5$. Students who were perfectly calibrated (e.g. 4 confident out of 4 correct) have Sensitivity that approaches 1 (e.g., $4.01/(.01+4.01)$ or $.998$) and those who were not accurately calibrated on any questions would have Sensitivity that was just over zero (e.g., $.01/(.01+4.01)$ or $.002$).

Demographics. Demographic information was provided to the author by the participating school districts. This information included student gender, grade-level, ethnicity (categorized in analyses to represent the largest groups: Hispanic, Asian, White, and Other), English Language Learner (ELL) status, and free/reduced lunch eligibility as a matter of socioeconomic status.

Analysis. As a first step, these data were analyzed in a way typical to calibration data: examining zero-order correlations between accuracy and calibration. Multiple regression analyses were then conducted using a dispositional framework to examine the associations between calibration and accuracy controlling for other observed student characteristics. Sensitivity and specificity were included together in the model to represent correct identifications of both confidence and uncertainty (see Schraw et al., 2013). To view the association between calibration and average improvement from pre to posttest, a model was estimated also controlling for pretest accuracy. In these single student-level models, each student's pretest means for accuracy, Sensitivity, and Specificity were calculated as was each student's mean posttest accuracy (as outcome). The final single student-level model is represented by the following equation:

$$\overline{\text{PosttestAcc}}_i = \beta_0 + \beta_1 \overline{\text{Sensitivity}}_i + \beta_2 \overline{\text{Specificity}}_i + \beta_3 \overline{\text{PretestAcc}}_i + \beta_3 \overline{\text{Covariates}}_i + r_i \quad (1)$$

To address the dynamic nature of calibration and its role within the model of conscious regulation as presented in Figure 3.1, a random intercepts two-level hierarchical model with

objectives nested within students was analyzed to determine whether student calibration at pretest (Sensitivity and Specificity) was associated with gains from pre to posttest. To isolate within-student effects and to eliminate bias from unobserved student characteristics (see Allison, 2005; Hofmann & Gavin, 1998; Park, 2008), group-mean centering (around each student's mean) was used for Level 1 predictors. In this way, the question of whether the same student made greater gains during objectives when he/she was better calibrated could be answered. Unchanging student characteristics were entered as covariates at Level 2 along with student means for pretest accuracy and calibration.

Level 1

$$\text{PosttestAcc}_{ti} = \beta_{0i} + \beta_{1i}(\text{Sensitivity}_{ti} - \overline{\text{Sensitivity}_{ti}}) + \beta_{2i}(\text{Specificity}_{ti} - \overline{\text{Specificity}_{ti}}) + \beta_{3i}(\text{PretestAcc}_{ti} - \overline{\text{PretestAcc}_{ti}}) + r_{ti} \quad (2)$$

Level 2

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \gamma_{01} \overline{\text{Sensitivity}_i} + \gamma_{02} \overline{\text{Specificity}_i} + \gamma_{03} \overline{\text{PretestAcc}_i} + \gamma_{04} \text{StudentCovariates}_i + u_i \\ \beta_{1i} &= \gamma_{10}; \beta_{2i} = \gamma_{20}; \beta_{3i} = \gamma_{30}; \beta_{4i} = \gamma_{40} \end{aligned} \quad (3)$$

The level one predictors are represented by formula (2), in which individual student mean (e.g., $\overline{\text{Sensitivity}_i}$) is subtracted from the student's score at that time-point (e.g., Sensitivity_{ti}) to arrive at the student group-mean centered value for that variable. Posttest accuracy is a function of the student intercept, these predictors, and a time-varying student error. The level two or student-level predictors are represented within formula (3) above. In this formula, the student-level intercept is a function of the grand intercept, the means of the calibration and pretest accuracy Xs, the non-time-varying student-level characteristics (gender, ethnicity, ELL, free/reduced lunch, grade-level), and a student error.

In this way, the within student effect for Sensitivity (β_w) is represented by γ_{01} and the between student effect (β_b) is represented by γ_{10} . The effect for the individual at level two is the difference between β_w and β_b —this “compositional” effect is the extent to which the student

effect remains once the individual quiz effect is controlled (see Raudenbush & Bryk, 2002). However, it is worth note that because this compositional effect does not control for lurking variables unique to the person, it should not be interpreted as definitive (see Allison, 2005).

To test for these differences between β_w and β_b , Wald post-estimation tests were conducted to compare within and between student coefficients for Sensitivity, Specificity, and pretest accuracy. Differences, if statistically significant, were quantified and expressed as differences in standardized effect sizes. All standardized effect sizes were calculated using the relevant level-specific standard deviation for each variable using the formula: $(B \cdot SD_X) / SD_Y$.

To answer the second research question regarding grade-level differences in calibration, analyses of variance were first conducted with student-level data using student mean levels of Sensitivity, Specificity, and accuracy. Tukey post-hoc tests were run to illuminate the results. Regardless of mean-level differences in calibration between students of different grade-levels, students at different developmental stages may use metacognitive information differently, resulting in differential associations between calibration and achievement gain depending on grade. To investigate these grade-level moderators of the association between calibration and quiz gains, a final model included interactions between grade-level and calibration measures.

Results

(1) Do students make greater pre to posttest gains when better calibrated at pretest?

The means and standard deviations of each measure of calibration along with pre and posttest accuracy are presented in Table 3.2, divided by grade. The top half of the table presents descriptive statistics at the observation level and the bottom half presents them at the student level. On average, students are more accurate in their posttest answers than their pretest answers

and are better able to correctly identify when they are correct (Sensitivity) than they are able to correctly identify when they may be incorrect (Specificity).

{Insert Table 3.2}

Variation in calibration is divided by the within and between student levels. Random effects Analyses of Variance reveal that 25% of the variance in Sensitivity and 29% of the variance in Specificity is associated with the student. Similarly, 17% of variation in pretest accuracy is between students.

Zero-order correlations between calibration and accuracy are presented in Table 3.3. Full correlations between calibration, accuracy, and other student characteristics are presented in Appendix B, Table 2. As in Study 1, Sensitivity and Specificity had a strong inverse correlation. Sensitivity had moderate correlations with both pretest and posttest accuracy; Specificity had weaker correlations with both. The accuracy measures were strongly correlated with each other.

{Insert Table 3.3}

Table 3.4 presents the results from the student-level regressions of posttest accuracy on pretest accuracy and calibration. The models progress in a step-wise fashion from a model that does not control for observable student characteristics (Model 1) to one that has a full complement of demographic controls (Model 3). The second model is seen as an intermediate step as it controls partially for student characteristics (grade-level), but also for characteristics of the task, as students at different grades received different curricula and quiz questions. Adding all observed student covariates does little to change explained variance—Model 3's r-squared is only a .01 improvement upon Model 1. The regression coefficients for calibration and pretest accuracy also change little with the addition of student covariates. In the final model, there is a strong association between pretest and posttest accuracy: a one standard deviation increase in

pretest accuracy is associated with a 0.72 standard deviation increase in posttest accuracy (based on the full sample standard deviations of 0.16 for posttest and 0.14 for pretest). The association between calibration and posttest accuracy is much smaller: $d=0.09$ for Sensitivity and $d=0.08$ for Specificity. This is slightly stronger than the statistically significant demographic associations. Male students, for example, score 0.06 standard deviations lower than female students on the posttest.

{ Insert Table 3.4 }

The models presented in Table 3.4 view calibration as a dispositional characteristic of the student and relate each student's average level of calibration to their average level of pre to posttest gain within the ST Math curriculum. A further series of models were estimated to account for both the dynamic nature of calibration and to estimate associations between calibration and achievement gain net of unmeasured student characteristics. Based on a random effects Analysis of Variance, 22% of variance in the posttest was associated with student as the grouping variable. This was confirmed with the unconditional hierarchical model specifying student as the nesting variable. Proportion of variance between the two levels and descriptions of incremental model fit are provided in Appendix B, Table 3. Student-level covariates produce a small but statistically significant improvement in explaining variance in posttest accuracy (2.61%). The addition of the pretest accuracy variables at both levels resulted in a larger improvement over the unconditional model (29.60%), and, whereas addition of the calibration variables resulted in a statistically significant improvement over this model, the incremental improvement was small (0.62%).

The left half of Table 3.5 displays the results from the hierarchical regressions. Compared to the unconditional model, the full conditional model without interactions explained 84% of the

variance between students and 15% of the variance within students. Pretest accuracy was a strong predictor of posttest accuracy from quiz to quiz ($d=0.30$; effect sizes calculated using the level-specific standard deviations of predictor and outcome variables). The mean of student pretest accuracy at level 2 is similar in magnitude to the association as calculated with the one-level model ($d=0.74$)—this combines both the quiz and student levels, subtracting the within coefficient from the between coefficient resulted in a 0.44 contextual effect for student (see Raudenbush & Bryk, 2002). Post-estimation Wald tests revealed that this difference was statistically significant at the $p<.001$ level. Both calibration measures were statistically significant predictors of within student differences in pre- to post-test quiz gains (Sensitivity: $d=0.07$; Specificity: $d=0.02$), and these coefficients were different from each other ($p<.001$ based on post-estimation Wald test). As means, both were also statistically significant predictors at the student level of mean growth from pre- to post-test (Sensitivity: $d=0.09$; Specificity: $d=0.08$). The difference between level one and two Sensitivity was small ($d=0.02$) and not statistically significant ($p=.66$). For Specificity, post-estimation Wald tests did indicate a statistically significant difference between the level one and level two coefficients ($p=.001$), with a difference of $d=0.06$; this contextual effect was different from that of Sensitivity ($p<.001$).

{Insert Table 3.5}

(2) Does calibration and the benefit from calibration vary depending on student grade-level? As seen from the means of calibration at the bottom of Table 3.2, Sensitivity and Specificity appear relatively stable across grade-levels. However, with regards to Sensitivity, proportion confident when correct, there was a statistically significant effect of grade $F(3, 3,908)=14.78, p<.0001$. Post-hoc analyses revealed fourth graders had lower Sensitivity than second, third, and fifth graders. With regards to Specificity, proportion uncertain when not

correct, there was also a statistically significant effect of grade $F(3, 3,908)=8.02, p<.0001$. Post-hoc analyses revealed that fourth graders had higher Specificity than second and third graders ($ps<.0001$), but there was no statistically significant difference between any of the other grades ($ps>.17$). In general, fourth graders were more likely to express uncertainty. This could be due to the difference in content within ST Math, as each grade-level experienced different content. This could also be due to differences between grades outside of ST Math: school or district-wide factors could have influenced the grades differently. To investigate these potential differences, school means of fourth grade Sensitivity and Specificity were examined. One school district with three study schools emerged as a consistent outlier. Excluding this district, fourth graders still had lower Sensitivity and higher Specificity than the other grades.

There were also grade-level differences in pretest accuracy, $F(3, 3,908)=31.01, p<.0001$. Second graders were more accurate than fourth graders ($p<.0001$), and fourth graders were less accurate than all other grades ($ps<.0001$). There remained a statistically significant effect of grade at posttest, $F(3, 3,908)=17.61, p<.0001$. Second graders remained the most accurate, and this difference was statistically significant in comparing second graders to third and fourth graders ($p<.0001$). Differences between third, fourth, and fifth graders did not arise to the level of statistical significance ($ps>.05$).

The right half of Table 3.5 displays results from the 2-level model including grade-level interactions. The depths of the slopes for Sensitivity and Specificity for each grade level can be seen in Figures 3.2 and 3.3. The direction and relative magnitude of coefficients for Sensitivity and Specificity were similar to those within the model without grade-level interactions: Sensitivity was a stronger predictor than Specificity at the within student level, but Specificity had a larger student contextual effect. At the within student level, the associations between

Sensitivity and posttest accuracy for second, third, and fifth graders were different than those for the reference group, fourth graders ($p < .05$)—the coefficient for Sensitivity was larger for fourth graders than for the other grade-levels. The associations between Specificity and posttest accuracy were not different between fourth graders and second and fifth graders ($p > .05$); however, third graders had a smaller coefficient for Specificity than did fourth graders ($p = .04$) and also smaller coefficients than those found for second ($p = .01$) and fifth graders ($p = .02$). A post-estimation test revealed that the third grade coefficient for Specificity was not statistically significantly different from zero ($p = .62$). At level two, coefficients for both Sensitivity and Specificity were lower for second, third, and fifth graders than for fourth graders ($p < .05$). Post-estimation test revealed that estimates for second, third, and fifth graders did not differ, however ($p > .05$). Eliminating the outlier district resulted in little change. The relative strengths of the coefficients for each grade level were unchanged; however, the difference in slope for Sensitivity between second and fourth graders did not arise to levels of statistical significance ($p = .09$).

Replication

As a robustness check, a replication was run using data from students who were in the study schools and participated in ST Math during the 2011-2012 school year. Because of the design of the study, this sample included those students who were in the 2010 sample and did not age out or otherwise leave their schools. It also included students who were in grades that became treatment grades at the start of the 2011 school year. Table 4 in Appendix B displays demographic information on this new sample. Starting in 2011, all students in grades 2 to 5 in all study schools used ST Math, and so the grade distribution is more even than in 2010, where fourth graders dominated (compare with Table 3.1). Also of note is the ethnic makeup of this new sample: the sample is less Hispanic (74% vs. 85%) and more White (20% vs. 8%).

Appendix B, Table 5 presents the descriptive statistics on accuracy and calibration for this sample and indicates differences between this sample and the 2010 sample. There were differences in calibration and accuracy by grade between the samples, but all differences save one were smaller than .08 (as the exception, 2011 third graders had pretest accuracy that was .10 higher than that of 2010 third graders, $p < .0001$). Although small in magnitude, most differences were statistically significant; however, there was no clear trend as to which sample had higher values on the variables, either across a given grade or a particular variable—of the 12 statistically significant differences, the 2011 sample had higher values for seven.

As with the 2010 data, around 20% of the variance was associated with student as the grouping variable (see Appendix B, Table 6). As variables are added to the model, the model fit improved to levels of statistical significance, with the final model (before interactions) improving on the unconditional model by 27.31% (compare with 29.60% for the 2010 sample). Hierarchical regression results for the 2011 sample are presented on the left side of Table 7 in Appendix B. Effect sizes for pretest accuracy at both levels were within $d = .05$ of the 2010 model; calibration effect sizes were within $d = .01$ at level one and $d = .03$ at level two, with direction and relative strength of sensitivity and specificity comparable between the two years. Also replicating the 2010 analysis, there was no statistically significant contextual effect of Sensitivity, but a contextual effect for Specificity that was larger than the within-student effect.

Grade differences in calibration remained in the 2011 sample. There was a statistically significant effect of grade on mean Sensitivity $F(3, 6,087) = 89.49, p < .0001$ and Specificity $F(3, 6,087) = 50.84, p < .0001$. Fourth graders had statistically significantly lower Sensitivity than second and third graders, as in the 2010 sample, but did not differ from fifth graders. In this sample, fifth graders also had statistically significantly lower Sensitivity than did second and

third graders, and third graders had statistically significantly lower Sensitivity than did second graders. With regards to Specificity, there was no statistically significant difference between third, fourth, and fifth graders, but second graders had lower mean-levels of Specificity than all three other grades.

Replicating the 2010 sample, there was a statistically significant effect of grade on both pretest accuracy, $F(3, 6,087)=204.68, p<.0001$, and posttest accuracy, $F(3, 6,087)=130.25, p<.0001$. Post-estimation tests revealed all grade-level differences were statistically significant ($ps<.0001$, although the pattern of results did not mirror that found in the 2010 data). Specifically, second graders were not the most accurate in this sample. Third graders were the most accurate at both pre and posttest in 2011—it is of note that approximately half of the third graders in this sample were second graders in the previous sample. Fourth graders did remain the lowest-scoring grade, however.

The right half of Appendix B, Table 7 displays results from the model including grade-level interactions. Figures 2 and 3 in Appendix B display the regression slopes graphically. At the within student level, the size of the coefficient for Sensitivity for the reference group, fourth graders, was similar to that from the non-grade-level interaction model. There were no statistically significant differences between fourth, fifth, and second graders in the magnitude of this coefficient; however, the association between Sensitivity and posttest accuracy for third graders was weaker than that for fourth graders and second graders ($ps<.05$). The associations between Specificity and posttest accuracy for second and fifth graders was statistically significantly larger than that for fourth graders ($ps <.01$) and for third graders ($ps<.001$ from post-estimation Wald tests), but not different from each other ($p=.12$). For the reference group, fourth graders, the coefficient for Specificity was not statistically significantly different from

zero. At level two, coefficients for both Sensitivity and Specificity were lower for third graders than for fourth graders ($ps < .05$), and the coefficient for Sensitivity was higher for fifth graders than for all other grades ($ps < .05$). Post-estimation Wald test revealed that second, third, and fifth graders differed with respect to the size of the Specificity coefficient at level two: the association for fifth graders was strongest and third graders weakest.

Discussion

(1) Do students make greater pre to posttest gains when better calibrated at pretest?

Analyses conducted with multiple measures of calibration and performance across a year-long mathematics curriculum were able to disentangle the associations between calibration and other aspects of regulation in a manner not previously undertaken within the calibration research. With respect to the first question, when better calibrated at pretest, students *did* make greater gains from pre to posttest. These associations were replicated with an additional sample of students, but were small, with effect sizes less than one tenth of a standard deviation. This is in contrast to the larger effects of calibration reported in previous research, mostly using between-person comparisons and zero-order correlations (e.g., Bol et al., 2010; Make, Shields, Wheeler, & Zacchilli, 2005; Ots, 2012; Soku & Quereshi, 2004).

{Insert Table 3.6}

Within the current study, effect sizes were sensitive to the mode of analysis: Table 3.6 compares effect sizes for Sensitivity and Specificity across methods. Sensitivity in particular appears to be inflated when using zero-order correlations, indicating that person characteristics associated with both Sensitivity and performance can bias results when analysis is conducted with this typical method. Estimates for Sensitivity changed little, however, between the one-level

model (Table 3.4) and the hierarchical model (Table 3.5)—it may be that Sensitivity's association with pretest accuracy is the relation causing the greatest bias—it was eliminated in both regression models by the addition of a pretest covariate.

Sensitivity and Specificity responded differently to mode of analysis and also had statistically significantly different associations with performance gains in the hierarchical model. Within student, the effect size for Sensitivity was more than three times as large as that for Specificity. It is a common feature of metacognitive and SRL models that students who are aware of what they don't know will engage in behaviors to direct and control learning, for example, students in ST Math could adjust attention, replay puzzles, or seek help (see Efklides, 2008; Nelson & Narens, 1990, Zimmerman, 2008). The small effect size for Specificity suggests that this may not be the case, at least when Specificity is measured at the Task x Person level, where student and task characteristics interact. It appears more important at this level that students are confident when they are correct. It is less clear what control mechanism is at play here. It may be that very task-specific self-efficacy is driving this association. The positive association between self-efficacy and performance is a consistent finding in the motivation literature (e.g., Bandura, 1997; Pajares, 1996).

Sensitivity and Specificity also differed in their associations with performance gains at the Person level. The level two effect for Sensitivity was no different than the within-person effect, but the contextual effect for Specificity was statistically significantly different from zero and three times as large as the within-person effect. This could indicate that there is some stable metacognitive monitoring trait that assists students in making and/or acting upon determinations of uncertainty. The kind of tendency toward deliberate thinking as noted in Winne (1995) may be

a candidate. However, as noted in Allison (2005), this estimate of the contextual effect may be biased by other stable characteristics of the student and should be interpreted with caution.

(2) Does calibration and the benefit from calibration vary depending on student grade-level? Looking at mean-levels of calibration across the grade-levels, it does not appear that the data supported the hypothesis that calibration accuracy improves with development in elementary-aged students (see Pressley et al., 1987). This could have been due to idiosyncrasies in the sample, especially within the 2010 data. The 2011 replication results differed. Within this sample, the younger students had higher Sensitivity, perhaps indicating greater confidence overall. Second graders had lower Specificity than the other grade-levels, indicating that they were not as accurate in identifying uncertainty for those questions they did not get correct. This is in keeping with prior research on age-related differences in monitoring, especially in complex or unfamiliar tasks (de Bruin & van Gog, 2012). Of all the grade-levels, the second graders would have been least familiar with multiple choice math tests as seen within ST Math; this may have driven their lower levels of Specificity.

Increasingly large associations between calibration and performance across the grade-levels would have indicated that older students used information from metacognitive monitoring judgments to differently engage control processes and subsequently make larger gains from pre to posttest. As with the analysis of mean-levels of calibration, no clear developmental picture emerged within the 2010 data. Unlike the mean-level analysis, the 2011 data did not appear to paint a clear picture either: there were no statistically significant interactions with respect to Sensitivity, and the higher coefficients for Specificity for both second and fifth graders as compared with third and fourth graders did not support the hypothesized developmental improvements.

Limitations. Using real-world data collected as part of an actual classroom mathematics activity offered some advantages over previous calibration research, but also provided some challenges. Because math content differed by grade-level, I was unable to disentangle grade and content—complexity and familiarity of test content may have played a large role in both the accuracy of calibration judgments and their use. The failure to replicate developmental differences across samples also reduces confidence in the results for the second research question. Additional replications, carefully adjusting content and difficulty differences between grade-levels, may offer stronger evidence for the presence of developmental patterns of overall calibration accuracy and use.

Although this study makes a substantial improvement in isolating calibration from the other aspects of conscious regulation as illustrated in Figure 3.1, even the within-person associations may be biased by other variables at the Task x Person level. For example, the effect of metacognitive knowledge about math tasks within ST Math would have been removed from the model at the Task x Person level; however, the difference between what a student knows about shape problems and what a student knows about fraction problems could be picked up in the calibration estimates between objectives, biasing the results. Using content that is more closely related may solve this problem, but a complete solution may remain elusive: as young students make great leaps in their math learning across the school year, the nature of their interaction with the task is likely to evolve, fundamentally changing aspects at the Task x Person level even in tasks that appear similar.

Conclusion. This study demonstrates the role of calibration within the system of SRL, showing how differences in the accuracy of student metacognitive monitoring are related to differences in performance gains within an online mathematics curriculum. As a main

contribution, this study moves beyond a dispositional view of calibration to explore the dynamic nature of calibration as it varies within the same student, from task to task. For the elementary school students within this study, calibration emerged as a statistically significant predictor of improvement in performance from pre to posttest. Within students, the accurate identification of correct answers (Sensitivity) had larger associations with performance gains than did the accurate identification of uncertainty for incorrect answers (Specificity), although both did have small statistically significant coefficients. This finding was replicated with a second sample of students. Although it was hypothesized that older students might both make more accurate calibration judgments and/or use those judgments more effectively toward performance gains, the results of this study did not conclusively support either hypothesis. These results can help calibration researchers in education and other fields start to disentangle the unique contribution of metacognitive monitoring within SRL—better understanding how calibration works in this dynamic system can help support the improvement of calibration and the improvement of SRL, both thought foundational to many learning activities.

References

- Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. SAS Publishing.
- Bandura, Albert. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs N.J.: Prentice-Hall.
- Bandura, A. (1997). *Self-efficacy: the exercise of control*. Macmillan.
- Barnett, J. E., & Hixon, J. E. (1997). Effects of Grade Level and Subject on Student Test Score Predictions. *The Journal of Educational Research*, 90(3), 170–174.
doi:10.2307/27542087
- 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(5), 372–382. doi:10.1016/j.learninstruc.2009.03.002
- 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(2), 133–151. doi:10.2307/20152656
- 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(4), 269–290. doi:10.2307/20157403
- Bol, L., Riggs, R., Hacker, D. J., Dickerson, D.L., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education*, 21, 81-96.
- Chen, P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences*, 14(1), 77–90.

- De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252.
doi:10.1016/j.learninstruc.2012.01.003
- Desoete, A., & Roeyers, H. (2006). Metacognitive Macroevaluations in Mathematical Problem Solving. *Learning and Instruction, 16*(1), 12–25.
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126C*, 213–228. doi:10.1016/j.jecp.2014.04.001
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science, 12*(3), 83–87.
doi:10.1111/1467-8721.01235
- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287.
doi:10.1027/1016-9040.13.4.277
- Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist, 46*(1), 6–25.
doi:10.1080/00461520.2011.538645
- Efklides, A., & Vlachopoulos, S. P. (2012). Measurement of metacognitive knowledge of self, task, and strategies in mathematics. *European Journal of Psychological Assessment, 28*(3), 227–239. doi:10.1027/1015-5759/a000145
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. doi:10.1037/0003-066X.34.10.906

- Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling Uncertain and Acting on It During the Preschool Years: A Metacognitive Approach. *Child Development Perspectives, 7*(3), 160–165. doi:10.1111/cdep.12035
- Hacker, D. J., Dunlosky, J., & Graesser, A. C. (1998). *Metacognition in Educational Theory and Practice*. Routledge.
- Hofmann, D. A., & Gavin, M. B. (1998). Centering Decisions in Hierarchical Linear Models: Implications for Research in Organizations. *Journal of Management, 24*(5), 623–641. doi:10.1177/014920639802400504
- Howie, P., & Roebbers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*(7), 871–893. doi:10.1002/acp.1302
- Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the Performance of Business Students on Examinations. *Journal of Education for Business, 79*(4), 217–224. doi:10.3200/JOEB.79.4.217-224
- Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296–298. doi:10.1016/j.learninstruc.2012.01.002
- 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*(6), 1121–1134.
- Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 663–679. doi:10.1037/0278-7393.10.4.663

- Markman, E. M. (1977). Realizing That You Don't Understand: A Preliminary Investigation. *Child Development*, 48(3), 986–992. doi:10.2307/1128350
- Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition*, 38, 441–451. doi:10.3758/MC.38.4.441
- Nelson, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. In Gordon H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 26, pp. 125–173). Academic Press. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0079742108600535>
- Newman, R. S., & Wick, P. L. (1987). Effect of age, skill, and performance feedback on children's judgments of confidence. *Journal of Educational Psychology*, 79(2), 115–119. doi:<http://dx.doi.org/10.1037/0022-0663.79.2.115>
- 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(2), 159–179. doi:10.1007/s10409-006-9595-6
- Ots, A. (n.d.). Third graders' performance predictions: calibration deflections and academic success. *European Journal of Psychology of Education*, 1–15. doi:10.1007/s10212-012-0111-z
- Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The Role of Individual Differences in the Accuracy of Confidence Judgments. *The Journal of General Psychology*, 129(3), 257–299. doi:10.1080/00221300209602099
- Pajares, F. (1996). Self-Efficacy Beliefs in Academic Settings. *Review of Educational Research*, 66(4), 543–578. doi:10.3102/00346543066004543

- Park, H. S. (2008). Centering in Hierarchical Linear Modeling. *Communication Methods and Measures*, 2(4), 227–259. doi:10.1080/19312450802310466
- Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, 4(1), 3–31. doi:10.1007/s11409-008-9030-4
- Pintrich, P. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, 16(4), 385–407. doi:10.1007/s10648-004-0006-x
- Pressley, M., & Ghatala, E. S. (1990). Self-Regulated Learning: Monitoring Learning From Text. *Educational Psychologist*, 25(1), 19–33. doi:10.1207/s15326985ep2501_3
- Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology*, 43(1), 96–111. doi:10.1016/0022-0965(87)90053-1
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. SAGE.
- Rinne, L. F., & Mazocco, M. M. M. (2014). Knowing Right From Wrong In Mental Arithmetic Judgments: Calibration Of Confidence Predicts The Development Of Accuracy. *PLoS ONE*, 9(7), e98663. doi:10.1371/journal.pone.0098663
- Roebbers, C. M. (2002). Confidence judgments in children’s and adult’s event recall and suggestibility. *Developmental Psychology*, 38(6), 1052–1067. doi:10.1037/0012-1649.38.6.1052
- Saxe, G. B., & Sicilian, S. (1981). Children’s Interpretation of Their Counting Accuracy: A Developmental Analysis. *Child Development*, 52(4), 1330–1332. doi:10.2307/1129526

- Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language*, *51*(1), 71–79. doi:10.1016/j.jml.2004.03.004
- Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp. 224–257). Cambridge University Press.
- Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development*, *15*(2), 115–134. doi:10.1016/S0885-2014(00)00024-1
- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48–57. doi:10.1016/j.learninstruc.2012.08.007
- Stone, N. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437–475. doi:10.1023/A:1009084430926
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1024–1037. doi:10.1037/0278-7393.25.4.1024
- Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist*, *30*(4), 173–187. doi:10.1207/s15326985ep3004_2
- Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research*, *41*(6), 466–488.

- Zhao, Q., & Linderholm, T. (2008). Adult Metacomprehension: Judgment Processes and Accuracy Constraints. *Educational Psychology Review*, 20(2), 191–206.
doi:10.1007/s10648-008-9073-8
- Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, 11(4), 307–313. doi:10.1016/0361-476X(86)90027-5
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329–339. doi:http://dx.doi.org/10.1037/0022-0663.81.3.329
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166 –183. doi:10.3102/0002831207312909

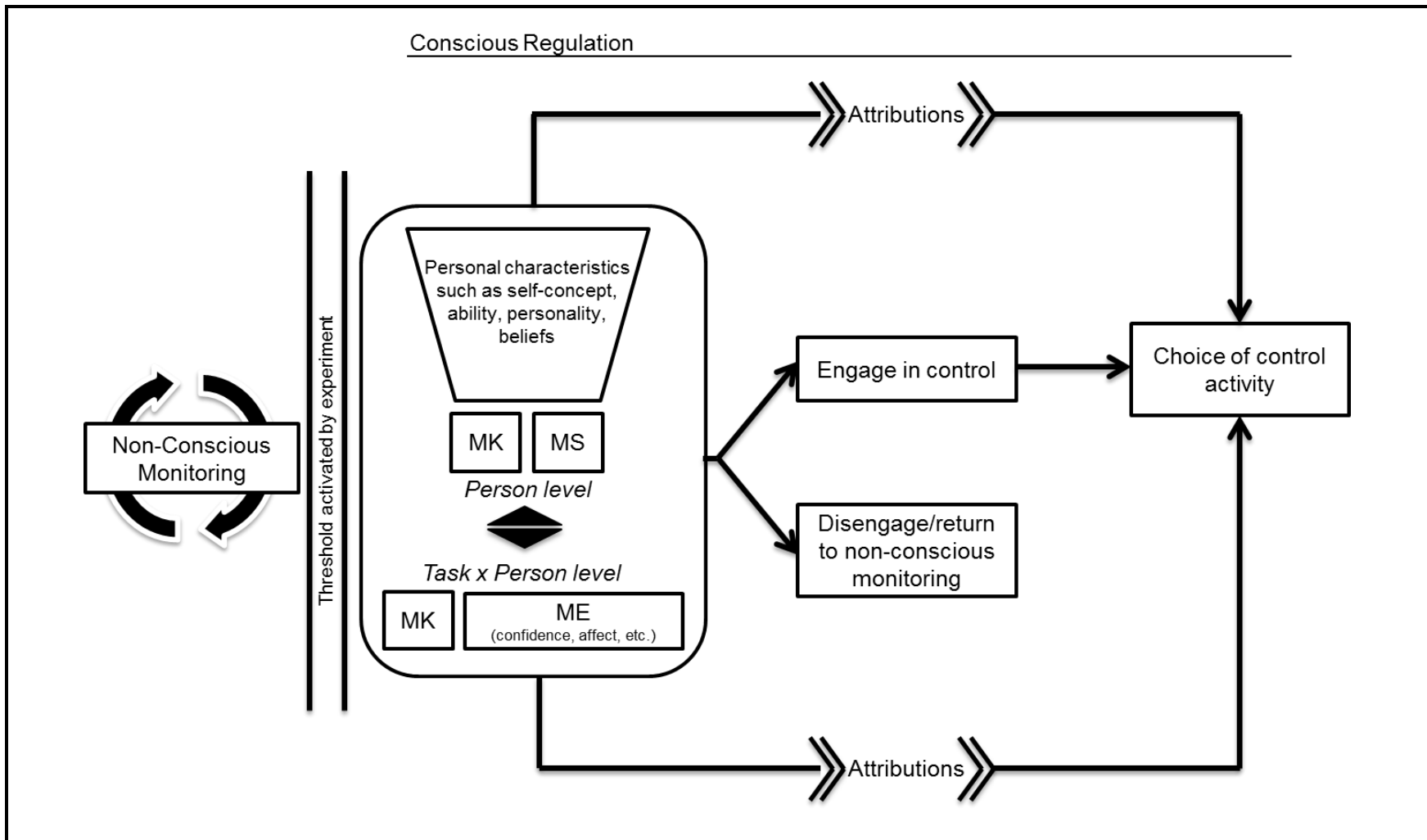


Figure 3.1. Model of conscious regulation. Flow chart illustrating monitoring and control after a request to make metacognitive judgments brings the learner into conscious regulation. Metacognition represented within as Metacognitive Knowledge (MK), Metacognitive Experiences (ME)—including confidence judgments, and Metacognitive Skills (MS).

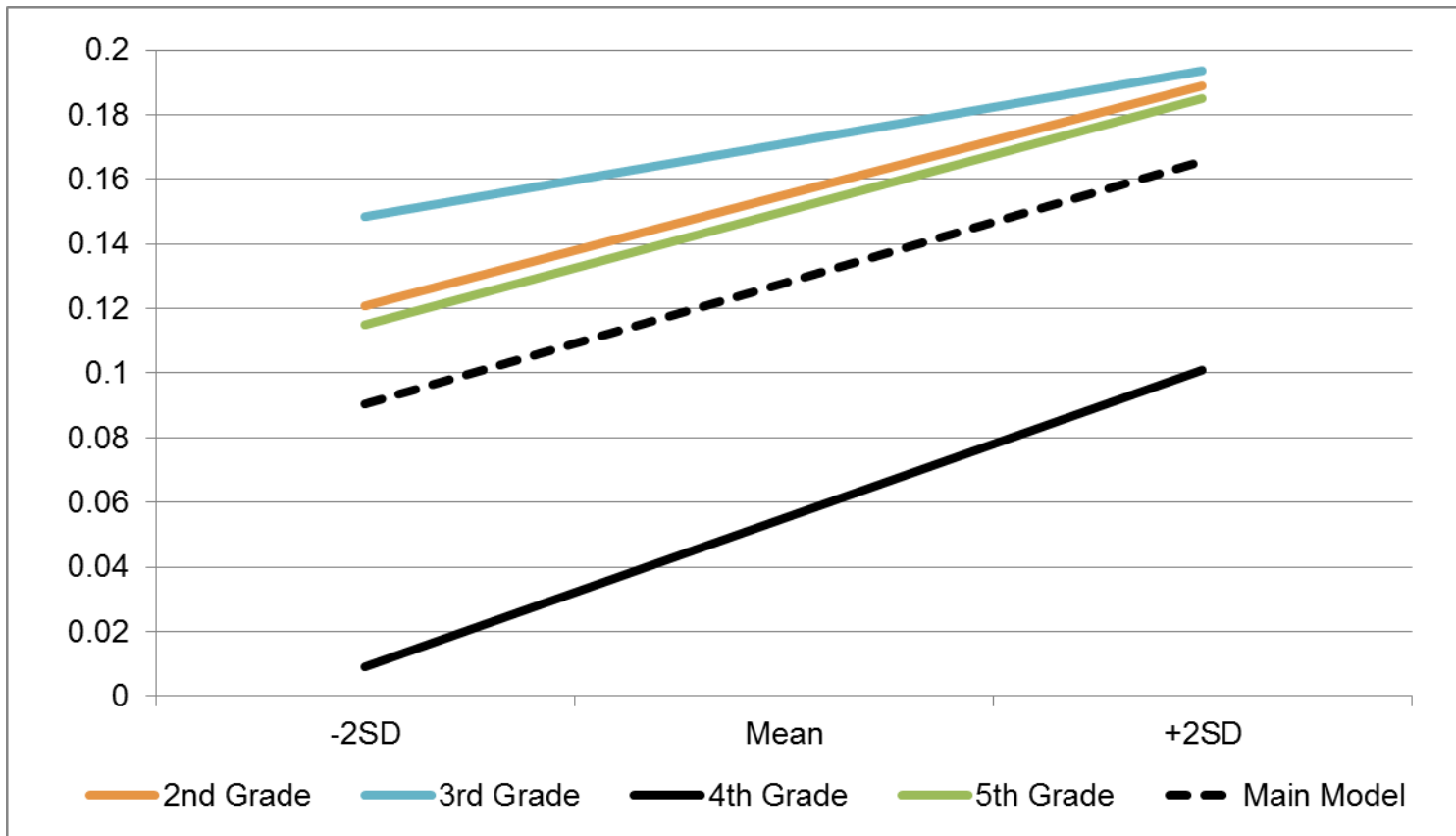


Figure 3.2. Level 1 slopes for Sensitivity by grade-level from interaction model on right-hand side of Table 3.5 compared with slope from non-interaction model (Main Model, dashed line) on left-hand side of Table 3.5.

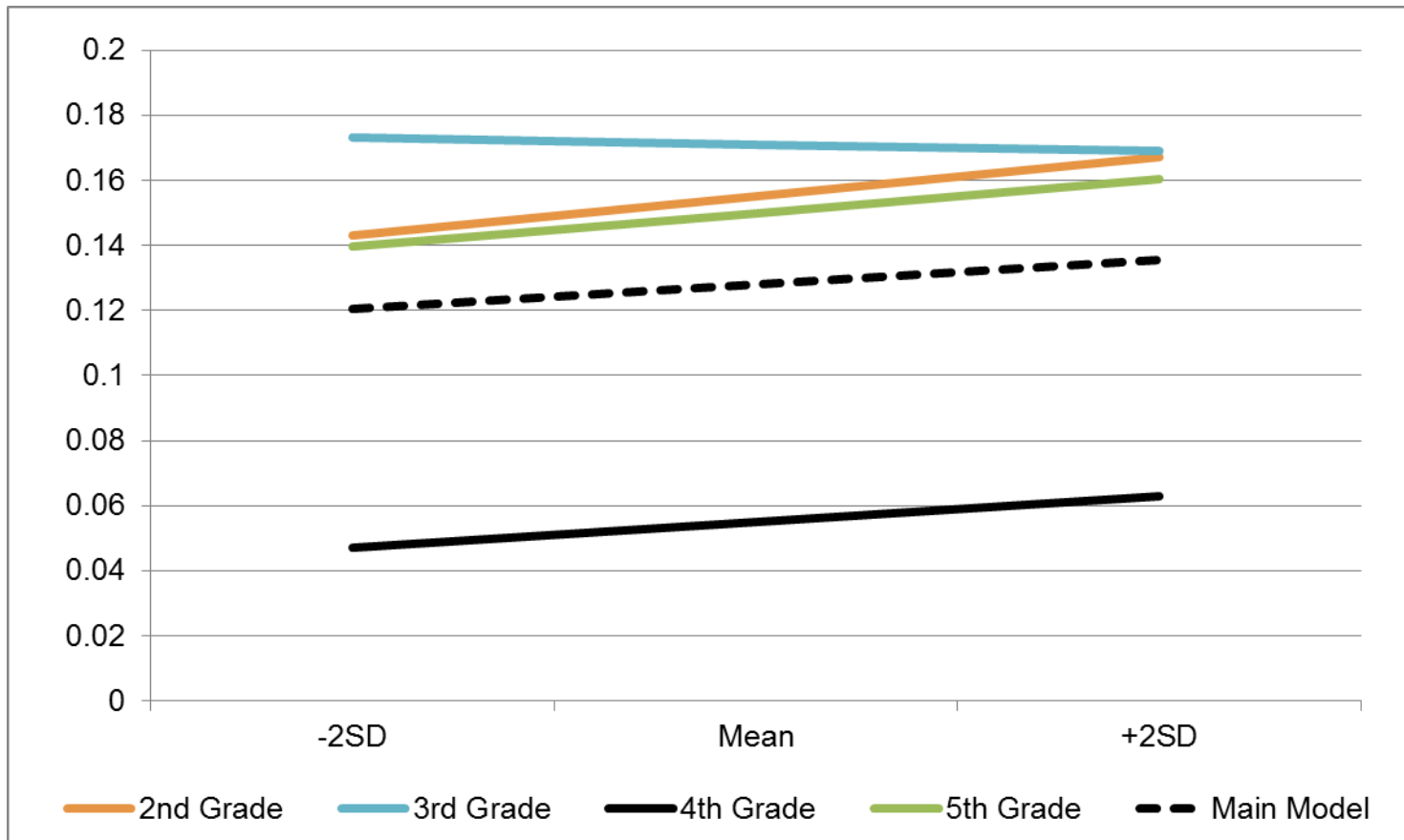


Figure 3.3. Level 1 slopes for Specificity by grade-level from interaction model on right-hand side of Table 3.5 compared with slope from non-interaction model (Main Model, dashed line) on left-hand side of Table 3.5

Table 3.1

Grade & Demographic Information of Study Students

	Total Sample		Analysis Sample	
	Percent	Count	Percent	Count
Grade 2	22%	4,137	21%	3,912
Grade 3	19%	4,137	19%	3,912
Grade 4	37%	4,137	37%	3,912
Grade 5	23%	4,137	22%	3,912
Male	52%	4,006	52%	3,912
Asian	3%	4,006	3%	3,912
Hispanic	85%	4,006	85%	3,912
White	8%	4,006	9%	3,912
Other Ethnicity	3%	4,006	3%	3,912
English Language Learner	66%	4,005	66%	3,912
Nat'l Free/Reduced Lunch	80%	4,006	80%	3,912
	N	4,137		3,912

Note. Total Sample includes all students in second through fifth grade in the study schools who began at least one objective within ST Math. The analysis sample is limited to those students who had complete demographic information and completed at least two complete objectives (pre and posttest).

Table 3.2
Quiz Accuracy and Calibration Measures, by Grade

Observation/Objective-Level Quiz Descriptives (N=56,962)								
	Grade 2		Grade 3		Grade 4		Grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest Accuracy	0.61	0.29	0.61	0.28	0.56	0.30	0.58	0.28
Pretest Sensitivity	0.82	0.31	0.84	0.29	0.78	0.34	0.82	0.31
Pretest Specificity	0.33	0.37	0.32	0.37	0.37	0.38	0.34	0.38
Posttest Accuracy	0.69	0.28	0.70	0.27	0.67	0.30	0.70	0.27
N (Observations)	12,935		11,146		21,263		11,618	

Student-Level Quiz Descriptive Statistics (N=3,912)								
	Grade 2		Grade 3		Grade 4		Grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest Accuracy	0.60	0.13	0.58	0.16	0.54	0.15	0.57	0.13
Pretest Sensitivity	0.82	0.18	0.82	0.17	0.78	0.19	0.81	0.17
Pretest Specificity	0.33	0.22	0.32	0.22	0.36	0.23	0.35	0.22
Posttest Accuracy	0.68	0.14	0.66	0.16	0.64	0.16	0.68	0.15
N (Students)	836		749		1,453		874	

Note. Data from analysis sample presented with objective data nested within students. Curricular and quiz content differs across grades.

Table 3.3
Correlations between Calibration and Accuracy Measures

N=3,912	Sensitivity	Specificity	Pretest Acc	Posttest Acc
Sensitivity	1			
Specificity	-0.79	1		
Pretest Acc	0.29	0.08	1	
Posttest Acc	0.25	0.07	0.81	1

Note. All correlations statistically significant at the $p < .001$ level.

Table 3.4
Student-Level Regressions of Posttest Accuracy on Pretest Accuracy and Calibration

N=3,912	(1)	(2)	(3)
Sensitivity	0.09*** (0.02)	0.09*** (0.02)	0.08*** (0.02)
Specificity	0.07*** (0.01)	0.06*** (0.01)	0.06*** (0.01)
Pretest Accuracy	0.83*** (0.01)	0.84*** (0.01)	0.82*** (0.01)
Grade 2		-0.01 (0.004)	-0.01 (0.004)
Grade 3		-0.01** (0.004)	-0.01** (0.004)
Grade 5		0.02*** (0.004)	0.01*** (0.004)
Eng Language Learner			-0.01* (0.00)
Male			-0.01** (0.03)
Asian			0.01 (0.01)
White			-0.0001 (0.01)
Other Ethnicity			0.002 (0.01)
Free/Reduced Lunch			-0.01*** (0.004)
Constant	0.09*** (0.01)	0.10*** (0.01)	0.13*** (0.01)
R2	0.66	0.67	0.67

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients. Standard errors in parentheses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.

Table 3.5
Results from Hierarchical Regressions of Post-test Accuracy on Pre-test Accuracy, Calibration, & Covariates

<i>Fixed Parameters</i>	B	SE	<i>p</i>	B	SE	<i>p</i>
Level 1						
Sensitivity	0.075	0.004	<.0001	0.092	0.006	<.0001
Specificity	0.015	0.004	<.0001	0.016	0.006	0.004
Pretest Accuracy	0.329	0.004	<.0001	0.329	0.004	<.0001
GR2*Sensitivity				-0.024	0.010	0.020
GR3*Sensitivity				-0.047	0.012	<.0001
GR5*Sensitivity				-0.022	0.011	0.045
GR2*Specificity				0.008	0.009	0.397
GR3*Specificity				-0.020	0.010	0.038
GR5*Specificity				0.005	0.009	0.631
Level 2						
Sensitivity	0.082	0.015	<.0001	0.149	0.023	<.0001
Specificity	0.057	0.012	<.0001	0.115	0.018	<.0001
Pretest Accuracy	0.843	0.012	<.0001	0.841	0.012	<.0001
Grade 2	-0.014	0.004	<.0001	0.100	0.035	0.004
Grade 3	-0.019	0.004	<.0001	0.116	0.035	0.001
Grade 5	0.011	0.004	0.003	0.095	0.036	0.008
ELL	-0.007	0.003	0.031	-0.007	0.003	0.032
Male	-0.008	0.003	0.005	-0.007	0.003	0.005
Asian	0.004	0.007	0.596	0.002	0.007	0.764
White	-0.001	0.005	0.911	0.000	0.005	0.961
Other Ethnic	0.009	0.007	0.237	0.008	0.007	0.254
Free/Reduced Lunch	-0.013	0.004	<.0001	-0.013	0.004	0.001
GR2*Sensitivity				-0.105	0.033	0.001
GR3*Sensitivity				-0.125	0.034	<.0001
GR5*Sensitivity				-0.072	0.034	0.037
GR2*Specificity				-0.085	0.027	0.001
GR3*Specificity				-0.099	0.027	<.0001
GR5*Specificity				-0.078	0.028	0.005
Intercept	0.128	0.014	<.0001	0.055	0.022	0.014
<i>Random Parameters</i>						
Between	0.003	0.0001		0.002	0.0001	
Residual	0.054	0.0003		0.054	0.0003	
<i>% Variance Explained</i>						
L2	0.832			0.888		
L1	0.148			0.148		

Note. Unstandardized regression coefficients. Level 1 variables are group-mean centered around student means. Level 2 quiz variables represent student means. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.

Table 3.6

Comparison of Effect Sizes for 2010 Data across Analysis Methods

	Zero-Order Correlations	One-Level Model	Hierarchical Model
Sensitivity	.25	.09	.07
Specificity	.07	.08	.02

CHAPTER FOUR

Study Three: Changes in Calibration: In Response to Intervention and as Related to Changes in Achievement

As students interact with learning tasks, they must set goals, evaluate their progress toward those goals, and adjust their strategies (Zimmerman, 2008). This system of self-regulated learning (SRL) is considered a crucial element of a positive mathematics disposition (DeCorte, Verschaffel, & Op'T Eynde, 2000); student strength at SRL has been linked with positive academic and life outcomes (Pintrich & de Groot, 1990; Zimmerman & Kitsantas, 2014). Within SRL, as students evaluate their progress, they make determinations as to their success or failure of goal attainment (Efklides, 2008; Winne, 2004). The accuracy of these judgments is termed *calibration*, and is seen as foundational to SRL and to learning activities in general (Alexander, 2013). Calibration itself has been shown to have a positive relation with achievement (Stone, 2000)—Study 2 within this dissertation demonstrated that accurate student judgments of confidence and uncertainty were associated with learning gains. Acknowledging the potential for calibration to improve SRL and performance, this study expands upon the work of the prior dissertation studies to investigate the malleability of calibration. Within, I evaluate the effects of a program to improve calibration in elementary mathematics students, reporting changes on multiple measures of calibration and on the relation between changes in calibration and changes in achievement.

Improving Calibration through Intervention

Much of the prior work on improving calibration has been conducted with research studies of college students in classes involving the acquisition of knowledge about a field (e.g., psychology). Work by Bol and colleagues (2001, 2005) focuses on the effect of practice tests to

improve student calibration. In one of the first studies to take place in an authentic education setting, Bol and Hacker (2001) compared two sections of the same research methods course, one that incorporated practice tests into student review sessions and one that had a standard review session without practice tests. Within both groups, calibration accuracy was associated with higher achievement, but surprisingly, the group who received practice tests was more poorly calibrated on midterm multiple-choice items. Qualitative information was collected on causal attributions, and taking these into consideration, the authors theorized that the treatment group focused too narrowly on the material covered in practice tests. No formal feedback was given on the practice tests, however, which Bol and Hacker noted as a possible explanation for the failure to see improvement in calibration. In another study, Bol, Hacker, O'Shea, and Allen (2005) similarly looked for causal attributions to describe lack of improvement in calibration accuracy after students practiced calibration with five online quizzes before the final exam. In this study, practice seemed to increase postdiction accuracy, but had no effect on prediction accuracy, a finding the authors posit may have to do with the measure of calibration obtained: student ratings of percentage of items answered correctly. Additionally, the authors reported that the calibration trajectory for the five quizzes was not the expected upward trending line—the failure to find such a result may have been due to the course material. As is typical in many college classes, quizzes likely did not cover content that built upon prior material; the calibration measures may have covered largely unrelated topics. Researchers have discussed the difficulty of transferring skills in calibration across items or tasks that may be viewed as unrelated (e.g., Keren, 1991). Studying trajectories of calibration is useful, as it can give insight into whether calibration is changeable (Greene & Azevedo, 2007), but in order to best estimate this change, trajectories

might need to include measures from more closely related tasks, such as those found in many mathematics classes (see Clements & Sarama, 2009).

Other studies also use test practice and review to improve calibration, but include additional elements to focus students on their performance and/or calibration. In a small experimental study, Labuhn, Zimmerman, and Hasselhorn (2010) assessed the effect of different feedback conditions on calibration and performance. After a brief lesson on order of operations, students were assigned to intervention conditions. All students were given an opportunity to practice the skill; those in the feedback conditions were given feedback either on their own performance relative to the maximum possible correct or on their own performance and the performance of others who had completed the same task. Feedback was presented graphically after each practice problem. Labuhn and colleagues found that students in both feedback conditions displayed better calibration on a posttest task than did the non-feedback group. Although the authors did not find an effect of the intervention on posttest performance for the entire group, they did find that students identified as poor performers at pretest improved both in their calibration and in their performance if they received the feedback directing them toward social comparisons. Open-ended questions of participants revealed that those who received feedback were more aware of the process of self-evaluation. The authors distinguished their study from prior work finding no benefits of feedback (e.g., Hacker, Bol, Horgon, & Rakow, 2000; Schraw, Potenza, & Nebelsick-Gullet, 1993), noting that the graphical element of their feedback condition may have induced reflection more readily than the feedback in the previous studies. This study can be distinguished from others in additional ways: Labuhn and colleagues' (2010) study was with elementary-aged students, whereas the Hacker et al. (2000) and Schraw et al. (1993) studies they cited were both conducted with undergrads. Additionally, the

unsuccessful studies only provided feedback in the form of available answers—they did not specifically call attention to the accuracy of each student's answer and calibration. Hacker and colleagues did provide students an opportunity to monitor their own calibration and instruction directing them to do so, but this may be different than specifically pointing out each individual's incorrect answers and flaws in calibration.

Huff and Nietfeld (2009), also in a study with fifth graders, did find positive effects on calibration from an intervention that provided students with correct answers for practice tasks without directly providing students with feedback on their individual answers. However, both of the intervention conditions included extensive direction to monitor comprehension over the course of 12 days of practice with passage reading and comprehension questions, including task-specific instruction and cues. All treatment and control groups improved their reading comprehension from pre to posttest—the groups that improved their calibration did not experience larger gains in reading performance. The authors noted that this may have been due to their use of a standardized measure of reading comprehension and to the brief nature of the intervention. Improving performance through interventions to improve calibration remains an elusive goal. One of the few studies to show a connection between calibration improvement and achievement involved a semester-long training of college students (Nietfeld, Cao, & Osbourne, 2006). This training combined feedback *and* explicit instruction on calibration strategies, directing the students to reflect on their calibration accuracy.

In their review of calibration research, Hacker, Bol, and Keener (2008) noted that across calibration training studies, feedback alone has failed to consistently improve calibration accuracy. However, much of the feedback studied was passive feedback (providing students with correct answers and leaving it to them to reflect). Although Labuhn and colleagues (2010)

attribute the success of their intervention to the graphical component, it is possible that giving students direct individual feedback, specific to them, also distinguished this study. Regardless of whether feedback (passive or direct) is *sufficient* to improve calibration accuracy, it is likely that feedback is a *necessary* condition for calibration improvement—feedback provides the means through which students can be directed to reflect on their calibration. Hattie and Timperley (2007) provide support for this assertion, noting that feedback on self-regulatory processes can help students develop their own error-detection skills (i.e., become better calibrated) and can "have major influences on self-efficacy, self-regulatory proficiencies, and self-beliefs about students as learners" (p. 90).

Although not directly related to calibration training, Dignath and Buttner's (2008) meta-analysis of interventions to foster SRL advocates for the use of feedback with elementary-aged children. The authors stress the complicated nature of metacognitive interventions, and note that younger children, who are still developing their metacognitive knowledge, may benefit from more scaffolds, including direct feedback on their SRL practices. The necessity of feedback to calibration interventions presents logistic challenges: the frequency and specificity of feedback required to improve calibration in elementary children may be too taxing for most elementary school teachers. Teacher-provided feedback presents an additional challenge: students may ignore it by augmenting or discounting (see Crocker, Voelk, Testa & Major, 1999; Hoyt, Aguilar, Kaiser, Blascovich, & Lee, 2007, noting that minority students may augment or discount feedback when given by others who know their race). Additionally, some participants in previous studies have specifically noted changing their predictions in an attempt to influence or protect against resulting negative self-concept (Dembo & Jakubowski, 2003, cited in Bol et al., 2005; see also Hattie & Timperley, 2007). Computer-provided feedback may present a viable

alternative for calibration training: computer feedback can efficiently and individually reach multiple students and provide objective information in an environment protected from negative social ramifications. Although there is little prior research on computer-assisted calibration training, prior literature on computer-provided feedback and educational technology interventions for SRL in general may prove illuminating.

Computer Provision of Feedback

Feedback is an extremely powerful tool for education, but it is rarely used effectively by teachers (Hattie & Timperley, 2007). Educational technology provides a means to improve the effective use of feedback. Computer-based learning environments can provide feedback frequently, individually, and without embarrassment. Hattie and Timperley (2007) claim that feedback has its strongest effects when it is in the "form of video-, audio-, or computer-assisted instructional feedback" (p. 84). Two recent studies on computer-provided feedback demonstrate its usefulness in math education.

Koedinger, McLaughlin, and Heffernan (2010) evaluated the formative assessment program ASSISTments using a quasi-experimental trial of its implementation with 1,240 seventh graders. ASSISTments provided individually directed feedback to students as they took computerized math assessments. The treatment group was compared to students at a similar school that was unable to implement ASSISTments because it did not have sufficient computers. Controlling for prior year's score, treatment students gained more than control students on a standardized measure of math achievement. The ASSISTments treatment included the provision of classroom progress information to teachers, which could have also been responsible for student gains if teachers changed classroom practices as a result of this information. A study of a similar system for homework has also shown benefits (Mendicino, Razzaq, & Heffernan, 2009).

In a randomized experiment, Mendicino and colleagues determined that a web-based homework system that provided immediate feedback to fifth grade students resulted in higher gains on classroom math assessments.

These two studies represent examples of the use of computer-provided feedback within the context of elementary and middle school mathematics. The use of computerized provision of feedback is not without caveats. The most beneficial feedback directs the student to think critically about her errors rather than merely drawing attention to inaccuracies (see Azevedo & Bernard, 1995). As described in Hacker et al. (2008) and as noted above, passive feedback or feedback on accuracy alone are unlikely to improve student calibration; however, both are essential in raising student awareness of the current status of their goal pursuits and allowing an avenue for implementation of metacognitive monitoring. Feedback on both the accuracy of an answer and on the metacognitive evaluation itself are important elements in directing students to reflect on their own SRL processes and are implicit components of the technology-based SRL programs described below.

Educational Technology and SRL

In the past decade, researchers have begun to investigate the affordances for SRL within computer learning environments (e.g., Dabbagh, & Kitsantas, 2005; Winters, Greene, & Costich, 2008). Dabbagh and Kitsantas (2005) explored the ways that web-based pedagogical tools within WebCT supported SRL practices of undergraduate students. They found that different types of tools supported different SRL processes such as goal-setting and self-monitoring through discussion, work sharing, and accessibility of standards. Winters and colleagues (2008) caution that although computer-based learning environments may offer tools to support SRL, many

students do not avail themselves of these tools, even if they report and perceive themselves as using the tools.

To support the use of these tools, metacognitive prompts in the form of questions embedded in the learning activity, can be used to direct student attention to their own SRL. These questions require students to plan, monitor, or evaluate their learning (Bannert & Mengelkamp, 2013). Through scaffolding student learning by directing them to self-regulate, such prompts can improve students' ability to learn complex material and can develop reflective habits of mind that contribute to SRL beyond the scaffolded topic (Azevedo, 2005; Bannert & Mengelkamp, 2013). SRL scaffolding and prompts within hypermedia environments have also been shown to increase student learning (Aleven & Koedinger, 2002; Azevedo & Hadwin, 2005). The success of metacognitive prompts is especially pronounced on transfer items that require greater depth of processing (Bannert & Mengelkamp, 2013). One example within the field of mathematics education is the metacognitive self-questioning program, IMPROVE (Kamarski & Gutman, 2006; Kamarski & Zeichner, 2001). Within IMPROVE, as students practice mathematical problem-solving, they are prompted to identify the type of problem they are working on, connect the problem to prior knowledge, use appropriate strategies to solve the problem, and reflect on their process. In a randomized trial, eleventh grade math students who received metacognitive questioning based on the IMPROVE system outperformed their peers assigned to a control condition that provided only feedback on problem accuracy (Kamarski & Zeichner, 2001). Kamarski and Gutman (2006) reported on another study with ninth grade math students assigned to an e-learning system either with or without IMPROVE. The students completed a five-week unit on linear functions within the e-learning environment and were

tested on procedural problems and "transfer" problems set in a real-life context. The IMPROVE students outperformed their peers on both types of problems.

A number of guidelines have been advanced regarding elements of successful metacognitive support programs. Bannert and Mengelkamp (2013) summarized these guidelines and noted that successful programs should (1) be integrated in domain-specific instruction, (2) explain the value of the support or instruction to students, and (3) allow sufficient training time. Additionally, Winters et al. (2008) noted that effects should be monitored once scaffolds are faded or removed and real-world achievement measures should be included to demonstrate how skills learned within the computer-learning environment can translate to outside learning or assessment situations.

Application of Prior Research and Formation of the Current Study

Azevedo (2007) defined computer-based "metacognitive tools" as "any technology-based environment that (to some degree) models, prompts, supports, and enhances a learner's self-regulatory processes," (p. 60) and includes monitoring behaviors, such as calibration, in his list of processes. The previously described studies of SRL in computer-based environments make use of such metacognitive tools to improve SRL in middle and high school students, where much of this research has been situated (see Azevedo & Hadwin, 2005). The success of programs like IMPROVE (Kamarski & Zeichner, 2001) is encouraging, but a similar integrated SRL program may not work for elementary-aged students, who are likely to require more direct training in the underlying components and strategies of SRL (see Bannert & Mengelkamp, 2013; Dignath & Buttner, 2008). As an initial step to improving SRL, the effectiveness of using a computer-based environment to train calibration alone can be investigated.

Elements of the online mathematics learning software, Spatial Temporal (ST) Math,

provide such training. Within ST Math, students progress through a grade-level mathematics curriculum to learn mathematical concepts by solving spatially-represented problems. In a randomized trial, students using an earlier version of ST Math had larger standardized math score gains than their control-group peers (Rutherford, et al., 2014; Schenke, Rutherford, & Farkas, 2014). Within the newest version of ST Math, 30+ quizzes throughout the year ask students to give confidence ratings about their answers by selecting a cheering (confident) or shrugging (not confident) icon for each question (Figure 4.1). This interface presents a novel way to approach the difficulty of training and assessing calibration with young children who struggle with traditional monitoring measures (see Huff & Nietfeld, 2009). Students are given graphical feedback about their confidence calibration after each quiz, allowing them to practice their own evaluative skills (Figure 4.1).

{ Insert Figure 4.1 }

This repeated practice at calibration with individual feedback on problem and metacognitive accuracy can guide students to reflect on this element of their SRL process. Although there is no direct instruction on monitoring or other SRL skills inherent in this system, feedback coupled with reflection has been shown to improve calibration accuracy (e.g., Nietfeld et al., 2006). To date, no studies of calibration training have been conducted in elementary school mathematics or over the course of an entire year. The ST Math calibration training may be especially effective due to the extended nature of the training and the practice within a domain such as mathematics, where skills build upon each other in a developmental progression (Clements & Sarama, 2009). The current study contributes to the research in calibration improvement by analyzing student change in calibration in response to practice and direct feedback and investigates the association between calibration changes and achievement gains

within the computer-based environment and on more traditional assessments. To this end, the following research questions are addressed:

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration?

(2) Is improvement in calibration accuracy linked to improvement in performance?

Method

Research Design. The current study uses data from an ongoing study of ST Math funded by an Institute of Education Sciences grant to a partnership between MIND Research Institute, the Orange County Department of Education, and researchers at the University of California, Irvine. Within the larger study, the effectiveness of ST Math was evaluated using a randomized control trial of 52 schools. The 52 elementary schools in the study included two cohorts with a staggered implementation design. This study will concentrate on Cohort 2 schools, which began implementing ST Math in the 2009-2010 school year. For these 18 schools, random assignment occurred during the summer of 2009. Nine schools were assigned to implement ST Math at grades 2-3 and not in grades 4-5 (Group A), and nine schools were assigned to implement ST Math at grades 4-5 and not in grades 2-3 (Group B). Although within schools the grades were split between treatment and control, the randomization occurred at the school level to either a second/third grade implementation or a fourth/fifth grade implementation. Thus, grades 2-3 of Group B served as controls for the treated grades 2-3 of Group A, and grades 4-5 of Group A served as controls for the treated grades 4-5 of Group B. The decision was made to assign all of a school's classrooms in a given grade as a group to either treatment or control to encourage fidelity to condition.

Each year in the study, schools had the option to add two treatment grades to provide multiple years of treatment to students as they progressed through elementary school. As a consequence, a third grade student who was assigned to a Group A (grades 2/3) school did not stop receiving ST Math in fourth grade so long as their school exercised the option to add additional grade-levels. Only one school elected to not add grades during their subsequent years in the study. The standard progression of treatment by grade for the 2007 to 2009 time-period is presented in Figure 4.2. The delayed treatment design utilized permitted variation in the number of years and grade-levels of those assigned to treatment, and supported equal engagement in the study by treatment and control teachers (e.g., Roschelle et al., 2010)—initial control teachers knew that their grade would receive the intervention within two years.

{Insert Figure 4.2}

Sample. Information on the overall ST Math sample and study school population are provided in Study 1. All Cohort 2 students who were in third or fourth grade during the 2011-2012 school year are included in this study (Study 3). By random assignment, roughly 50% of these students received ST Math in 2010-2011 as second or third graders (the Early Treatment Group, ETG) and roughly 50% started ST Math the following year, in the 2011-2012 school year (the Late Treatment Group, LTG). Of the 6,091 ST Math students in 2011 (identified in Study 2, above), 2,990 were in the appropriate grades for this study. The analysis sample for this study was limited to those students who stayed in the same treatment group for both years and who have valid standardized test achievement data for both years (N=2,625, 88% of the possible sample). The sample draws from 18 schools; distribution of the study students among the 18 schools range from 81 students to 248 students at a given school, with a mean number of 146 students at each school.

{Insert Table 4.1}

Table 4.1 displays the demographic information on the sample, divided between the early and late treatment groups. Although the schools were assigned to treatment group randomly, and each contained both treatment and control grade-levels at random assignment, there are some differences in demographic characteristics that varied with school X grade combinations. The ETG had more third graders and fewer fourth graders ($p=.022$), fewer Hispanic students and more students of all other ethnic groups ($ps<.05$), and fewer students eligible for free/reduced lunch ($p=.001$).

Measures.

ST Math quiz data. MIND Research Institute provided in-game quiz scores and calibration measures for each of the treatment students. Data included item-by-item quiz answers, accuracy, and confidence ratings. Each year included 23 pre and posttest quizzes, administered to students as they completed the ST Math curriculum. As students started a new objective, they took a pretest on the content within that objective and specified their confidence (sure or not sure) in each answer they gave. After the objective, they took a posttest, also selecting their confidence level.

As choice of calibration calculation may influence results, as recommended by Dunlosky and Thiede (2013), multiple measures of calibration were calculated. In keeping with Studies 1 and 2 of this dissertation and recommendations by Schraw, Kuch, and Gutierrez (2013), Sensitivity and Specificity were examined. To aid in comparability across other studies of calibration, three other commonly used measures were chosen: Simple Match, Gamma, and Discrimination (see Schraw et al., 2013). For all measures, higher scores indicate better calibration, but the measures may represent different types of calibration: Gamma and

Discrimination represent a student's ability to differentiate between items (relative calibration), Simple Match represents a student's absolute calibration, and Sensitivity and Specificity represent different processes of calibration for judgments of confidence and uncertainty (see Study 1; Kuch, 2012; Schraw et al., 2013). Because patterns of responses would have rendered some measures incalculable (see Study 1), the procedure from Study 2 was followed: .01 was added to each of the quadrants representing accuracy/confidence combinations before calibration measures were calculated.

Standardized Test Scores. Scores from the California Standards Test (CST), administered to all California students grades 2-11 in the spring of each year, were used to assess mastery of grade-level mathematics content. CSTs are criterion-referenced, standards-based assessments developed in alignment with the California Content Standards (California State Board of Education, 2010a). For the 2007-2008 test administration, the latest year for which this information is available, Cronbach's alphas in grade 2 and 3 CST mathematics were 0.93 and 0.94, respectively (ETS, 2008). Scale scores ranging from 150 to 600 were calculated by the state to allow for comparison between grade-levels and were provided to the IES study researchers by the participating school districts. These scale scores are necessary because tests are designed to assess each grade's standards and therefore differ between grades. Within grades, each year's test is based on the same core of standards, but contains different questions from the years prior. Across math and English/Language Arts (ELA) in all elementary grades, a scale score of 350 points indicates a student is considered by the state to be proficient in that subject's content-matter for that grade. In addition to specifying the 350 point proficiency cut-off, the state of California has designated math cutoff points for far below basic (scores less than approximately 240, depending on grade level), below basic (below 300), basic (300-350) and

advanced (above 400, with the exact value depending on grade level) (California State Board of Education, 2010b). ELA scores were used as covariate controls in most models.

Demographics. Gender, ethnicity, free/reduced lunch, and ELL status were reported by the school districts along with the CST data. Ethnicity was represented in the analyses by five groups: Hispanic, Asian, Black, White, and Other, to represent the largest ethnic groups within the sample. Reported English Language Learner (ELL) status was determined by schools as measured by the California English Language Development Test (California Department of Education, 2011). Eligibility for the national free and reduced lunch program is used as a measure of socioeconomic status.

Analyses.

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration? The randomly assigned variation in exposure to the training was used to answer this question. To answer whether there were differences in calibration after one year of ST Math calibration training, calibration scores from the 2011-2012 school year were regressed on treatment group. For the ETG, the start of 2011-2012 is one year after treatment, and for the LTG, it is before treatment (or just at the start of treatment). Measures of calibration were calculated from the 2011-2012 posttest quizzes to equalize familiarity with the ST Math interface and each specific objective arena. The first set of analyses used quiz calibration from the first objective encountered by the students. Because task-specific knowledge may have been affected by treatment group and may have effects on measures of calibration (see Study 2; Efklides, 2008), same-objective pretest accuracy was entered as a control variable. Grade-level (whether student was a fourth grader) was entered as a covariate as were demographic variables: gender, ethnicity,

ELL and free/reduced lunch status. To capture general mathematics knowledge and academic performance, pretest math CST scores were also entered into the model. Pretest ELA CST scores were included in the model to control for additional pretest academic characteristics. The OLS Regression equation for this analysis is represented as Equation (1). To account for nesting of students within schools, Huber-White clustered standard errors were used in this and other models.

$$\text{Calibration}_i = \beta_0 + \beta_1\text{ETG}_i + \beta_2\text{PretestAccuracy}_i + \beta_3\text{FourthGrade}_i + \beta_4\text{PriorAchievement}_i + \beta_5\text{DemographicVariables}_i + e_i \quad (1)$$

The model was run for each of the five calibration variables: Sensitivity, Specificity, Gamma, Simple Match, and Discrimination. For more stable measures of calibration, two aggregations of calibration were created for each calibration measure: a year-long aggregation for the 2011-2012 school year and an aggregation of the first three quizzes each student encountered. Due to the self-paced nature of the ST Math curriculum, these aggregated measures may include results from different objectives for each student. To accommodate this, dummy variables were included in these aggregated models as indications of which objectives contributed to the estimates and to compare students only to peers who had completed the same objectives. The year-long aggregated model also included a variable for the number of encountered objectives (1-23).

The prior year (2010-2011) CST scores were endogenous to the treatment for the ETG. To ascertain the effect of this endogeneity on regression results, a robustness check was performed using the 2009-2010 CST scores as math and ELA pretests. In this year, the ETG received ST Math, but did not receive the calibration intervention within ST Math, as it was only added to the version of the software introduced in the 2010-2011 school year. These analyses

were limited to fourth graders because third graders were in first grade in 2009-2010 and therefore did not take the CSTs.

Due to the nature of the provided calibration data, a posttest-only design was necessary. Although the schools were randomly assigned to study groups, it is possible that the two groups started with different levels of calibration. To test the likely magnitude of bias due to this limitation, calibration of the two groups (ETG and LTG) were compared from their first treatment year (2010-2011 for the ETG and 2011-2012 for the LTG). Only those students who were matched across 2010 and 2011 samples were included for the ETG (N=1,259, 99% of the main question sample). These students were in second and third grades in 2010—comparing them to the LTG sample in 2011 compares ETG second graders to LTG third graders and ETG third graders to LTG fourth graders. Because of this, grade could not be controlled as second or fourth grades would be perfectly collinear with treatment group. Nor could CST scores be included because second graders would not have pretest CST scores. The equation for this analysis is represented by Equation (2) below. As with the main analysis, three sets of equations were run using different samples of objective quizzes: the first objective only, the first three objectives, and all objectives from the first year of treatment (2010 for the ETG and 2011 for the LTG). An additional analysis was run using the entire-year aggregation of quizzes limiting the sample to only those students in fourth grade in 2011. In this analysis, pretest CST scores (from 2009-2010) could be controlled.

$$\text{Calibration}_i = \beta_0 + \beta_1\text{ETG}_i + \beta_2\text{PretestAccuracy}_i + \beta_3\text{DemographicVariables}_i + e_i \quad (2)$$

(2) Is improvement in calibration accuracy linked to improvement in performance?

Because of the timing of calibration and achievement measures, direct tests of the mediation of calibration on the relation between treatment and math achievement could not be conducted.

However, the association between ST Math calibration and math achievement, both in and outside of the software, can be explored with data from 2011-2012 for both the ETG and LTG students. This was done two ways: using gain across two objectives and using slopes of calibration to predict slopes of achievement.

For the first method, within each grade, two objectives were identified: one at the beginning of the curriculum (within the first five objectives) and one at the end of the curriculum (within the last five objectives that at least 50% of the students completed). Although all ST Math objectives cover on-grade mathematics curriculum in a similar manner (with spatial puzzles), there are some topics that may be more related than others. With this in mind, correlations between pretest accuracy for each objective were analyzed to determine which pair had the strongest correlation. It is for this pair that change scores were calculated: subtracting the early calibration from later calibration and early achievement from later achievement. The model is represented by equation (3) below and was estimated for each of the five measures of calibration separately. An additional series of models investigated the links between changes in calibration and changes in achievement outside of the ST Math curriculum, substituting the delta for mathematics CST score change for that of in-game math quiz accuracy.

$$\Delta\text{PosttestAccuracy}_i = \beta_0 + \beta_1\Delta\text{Calibration}_i + \beta_2\Delta\text{PretestAccuracy}_i + \beta_3\text{ETG}_i + \beta_4\text{FourthGrade}_i + \beta_5\text{DemographicVariables}_i + e_i \quad (3)$$

A second set of analyses used all available quiz information to calculate a slope for calibration improvement for each student, separately for each measure. This slope represents each child's calibration improvement over time as they progressed through the ST Math curriculum. The slopes were calculated by regressing calibration on ST Math objective quiz number (1 through 23). Similar slopes were calculated for posttest accuracy: regressing accuracy on ST Math objective quiz number. Once the slopes were calculated, growth of achievement was

regressed on growth in calibration, controlling for average pretest accuracy (as a measure of initial topic-specific knowledge), grade, treatment group, and demographic variables, along with the total number of objectives seen and a series of dummy variables to indicate which objectives contributed to the estimates. Equation (4) represents the model.

$$BPosttestAccuracy_i = \beta_0 + \beta_1 BCalibration_i + \beta_2 AvePretest_i + \beta_3 FourthGrade_i + \beta_4 ETG_i + \beta_5 DemographicVariables_i + \beta_6 TotalObjectives_i + \beta_7 DummyVariablesObjectives_i + e_i \quad (4)$$

This model was estimated for each of the five calibration measures. An additional set of models (equation 5) was estimated to explore how growth in calibration across the year related to achievement on the end-of-year (2012) mathematics CST score, controlling for the prior year's (2011) CST scores.

$$2012MathCST_i = \beta_0 + \beta_1 BCalibration_i + \beta_2 FourthGrade_i + \beta_3 ETG_i + \beta_4 DemographicVariables_i + \beta_5 TotalObjectives_i + \beta_6 DummyVariablesObjectives_i + \beta_7 2011CST_i + e_i \quad (5)$$

Results

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration?

Comparisons of mean-level differences between variables of interest and continuous covariates are presented in Table 4.2a (third grade) and 4.2b (fourth grade). Third graders who had already completed one year of the program completed more of ST Math—they saw, on average, one more objective than did the third graders who were new to ST Math ($p=.004$). However, using these unadjusted means, the ETG was less accurate in their quiz performance and in all measures of calibration ($ps<.001$) except Specificity: the ETG's average quiz pretest Specificity was higher than that of the LTG ($p=.005$). There were no statistically significant differences between the third grade ETG and LTG in mathematics standardized test scores, but there were differences between the two in ELA scores: the ETG group had lower ELA scores in 2011 ($p=.01$), and higher scores in 2012 ($p=.01$).

{Insert Table 4.2a}

{Insert Table 4.2b}

Among the fourth grade sample, the ETG and LTG did not differ in the number of objectives they completed ($p=.79$), nor in their quiz score accuracy ($ps>.05$). Similar to the third grade sample, however, the ETG had lower calibration scores on Sensitivity at both pre and posttest ($ps<.0001$), and on Gamma and Simple Match, but only at posttest ($ps=.02$). Like the third grade sample, the ETG had higher Specificity, at both pretest ($p<.0001$) and posttest ($p=.0004$). ELA CST scores were also different between the two groups: the ETG had lower scores for 2011 and 2012 ($ps=.004$).

Table 4.3 displays the correlations of calibration measures and their correlation with quiz accuracy and CST achievement test measures. As expected (see Study 1; Schraw et al., 2013), calibration measures were correlated. Sensitivity and Specificity showed weaker correlations with all other measures than they did with each other: they were strongly negatively correlated at both pre and posttest. All measures were also relatively stable from pre to posttest: same-measure correlations were above .5 except for Discrimination, which was just under at .49. Pretest Specificity was the only measure that did not show a statistically significant correlation with quiz pretest accuracy; however, posttest Specificity did show a small, statistically significant ($p<.01$) correlation with both pre and posttest accuracy. Similarly, Specificity showed the weakest correlations with CST scores. Quiz accuracy and CST scores had correlations around .5. ELA CST scores had stronger correlations with quiz accuracy than did math CST scores.

{Insert Table 4.3}

Although the groups were randomly assigned, unadjusted comparisons between the ETG and LTG may be biased by demographic or achievement characteristics that differed between the

groups. A series of regression equations controlled for these characteristics and also for task-specific knowledge (an important feature of the Task x Person level of metacognition, see Study 2; Efklides, 2008, 2011). The first set of regressions focused on the first objective encountered by students at the start of the 2011-2012 school year. This comparison was thought to capture the cleanest difference between treatment groups: the first objective would be before the LTG had significant opportunity to practice or receive feedback on calibration. For most students, this was an objective covering the topic of Place Value (see Appendix C, Tables 1a and 1b). Although it appeared that a large number of fourth graders encountered the Symmetry objective first, those who did not encounter it first, did not complete it at all, whereas 98% of students completed the Place Value objective. For this reason, the first analyses focused on Place Value and the 2,560 students (98% of the complete sample) who had valid data for this objective.

Table 4.4 displays the regression results of calibration on treatment group. The full results, including coefficients for covariates, are available in Appendix C. The ETG had lower calibration for every measure except for Specificity. Pretest accuracy also had statistically significant associations with each calibration measure except for Specificity. Of the covariates, math and ELA CST scores were positively related to calibration, but these relations only achieved statistical significance for four of five calibration measures for math and for three out of five for ELA. Fourth graders had higher calibration across all measures.

{ Insert Table 4.4 }

For a more stable measure of calibration, two additional sets of analyses were run on aggregations across quizzes. Table 4.5 presents the results for the regressions run using the first three quizzes and Table 4.6 presents results using all taken quizzes. The results for most calibration variables were largely in line with the regressions in Table 4.4—the ETG displayed

lower levels of calibration—Beta values were within .04 across samples. However, the effect of ETG on Specificity increased from $\beta=.02$ using the first objective to $\beta=.06$ using all objectives, gaining statistical significance ($p=.02$) in the aggregated analysis.

{ Insert Table 4.5 }

{ Insert Table 4.6 }

As a robustness check, the fourth grade sample was reanalyzed controlling for an earlier CST, taken at the end of the 2009-2010 school year, before either group was exposed to the calibration practice and feedback within ST Math. The results are available in Tables 5 through 7 in Appendix C. Beta coefficients are similar, with a negative effect of ETG seen for Sensitivity, Simple Match, Gamma, and Discrimination, and a positive effect of ETG seen for Specificity. The most notable change is within the model using the entire-year aggregation of calibration scores: effects for Simple Match, Gamma, and Discrimination lose statistical significance, whereas Sensitivity and Specificity have larger beta weights than those from models using the 2011 CST controls and including third graders.

The next series of analyses explored whether there were calibration differences between the ETG and LTG before treatment. To do this, the ETG's first year of treatment, 2010, was compared to the LTG's first year of treatment, 2011. Mean-level changes within the ETG from 2010 to 2011 were first examined. Of the 1,272 ETG students who were present in 2011, 1,259 (99%) of them had valid data for their first year of treatment, 2010. Table 4.7 compares calibration and accuracy variables across the years for this group. Those students who started in second grade increased on all calibration and accuracy measures from second to third grade—these differences were statistically significant ($ps<.05$) for all measures but Sensitivity. For students starting in third grade in 2010, differences between years did not arise to statistically

significant levels for posttest accuracy or any of the combined calibration measures (Simple Match, Gamma, Discrimination). However, pretest accuracy and Sensitivity were lower in 2011 than in 2010 ($p < .0001$) and Specificity was higher ($p < .0001$).

{Insert Table 4.7}

Results from regression analyses comparing the ETG at their year 1 (2010) to the LTG at their year 1 (2011) are presented in Appendix C, Tables 8 through 10. These results represent the association between assigned treatment group and pretest levels of calibration, before implementation of the experimental manipulation. Because some students in the ETG were in second grade in 2010, a CST pretest covariate could not be included. A robustness check on the final set of analyses (using the year-long aggregation) analyzes only those students who would be in fourth grade in 2010-2011, allowing for the control of the end-of-2010 CST scores for both the ETG and LTG (Appendix C, Table 11). Table 4.8 displays the beta coefficients from these pretest analyses (Yr 1 vs 1) and compares them with the beta coefficients from the after-treatment analyses (Yr 2 vs 1). The top half of the table displays these comparisons from the entire sample, and the bottom half from the fourth grade sample only, allowing for a CST pretest covariate exogenous to treatment.

{Insert Table 4.8}

For Simple Match, Gamma, and Discrimination in the entire sample, the ETG started with levels of calibration lower than those of the LTG. When the sample was limited to only fourth graders and the pretest CST was included, these differences did not rise to the level of statistical significance. After one year of treatment, these measures of calibration appeared to improve for the ETG, but remained lower than those of the LTG. The ETG also had lower levels of Sensitivity than did the LTG in their first year of ST Math, though these differences did not

attain statistical significance in the fourth grade sample with the CST pretest controlled. At year 2, however, both the full and limited samples showed lower Sensitivity for the ETG than for the LTG. Within both samples, Specificity was higher for the ETG before treatment, but did not reach levels that attained statistical significance. After one year of treatment, the ETG had higher Specificity than did the LTG in their first year of treatment, and these differences were statistically significant.

(2) Is improvement in calibration accuracy linked to improvement in performance?

For this question, two analyses were conducted: the first looking at gains in calibration between two related objectives and the second looking at the growth of calibration represented as a slope. For the first set of analyses, correlations between objectives were analyzed to choose the appropriate objective pair. These correlations, broken down by grade, are displayed in Appendix C, Table 12. Although almost all are statistically significant ($p < .05$), most are low (below .3). For third graders, Objective 14, Measurement, had relatively strong correlations with the first objectives, the strongest of which was with Objective 2, Ordering and Comparing Whole Numbers, $r(681) = .35$, $p < .0001$. On average, fourth grade correlations between objectives were weaker than those in third grade. However, the strongest pairing had a similar correlation to that found for the chosen third grade pair. This fourth grade pairing was between Objective 13, Decimal Operations and Money, and Objective 4, Whole Number Addition and Subtraction, $r(901) = .36$, $p < .0001$.

As not all students took all objectives, the sample for this analysis was limited to those students who had both the paired objectives, resulting in a reduced sample of 1,586 students (60% of the sample included for Question 1). Table 4.9 compares those students included in this sample to those students excluded. All differences between the samples attained statistical

significance except for the difference in the number of White students: White students made up 21% of both the included and excluded samples.

{Insert Table 4.9}

Table 4.10 displays descriptive statistics on the included measures of calibration and achievement, divided by grade, and provides the simple gain scores and standard deviations of these gains. From the early objective to the later objective, third graders made gains in pre and posttest accuracy and in all measures of calibration. Fourth graders declined in accuracy and all calibration measures except for Specificity. Both grades made gains in math CST scores, but only fourth graders made gains in ELA CST scores. There appears to be variance in these gains; the regression analyses for this question focus on whether the variation in calibration gains covaries with variation in posttest achievement gain.

{Insert Table 4.10}

The results of the regression analyses are displayed in Table 4.11. Only gain in Sensitivity, or proportion confident when correct, was associated with positive gain in posttest accuracy from the early quiz to the later quiz ($\beta=.07$, $p=.02$). Increases in Specificity, proportion uncertain when incorrect, were associated with decreases in posttest accuracy across the quiz pairs. Gain in the other measures of calibration did not have associations with posttest performance gain that attained statistical significance. Gain in pretest accuracy across the quiz pairs was associated with gain in posttest accuracy; being a member of the ETG was also associated with greater gain in posttest score. Coefficients for other covariates are shown in Appendix C (Table 13). Table 4.12 explores whether these gains in calibration are associated with gains in math CST scores from the end of the prior year (2010-2011) to the end of the studied ST Math year (2011-2012). No statistically significant relations emerged between calibration and achievement gain.

{Insert Table 4.11}

{Insert Table 4.12}

For the second set of analyses, growth in calibration and accuracy were represented by calculating slopes of pretest quiz calibration and of posttest quiz accuracy. Table 4.13 presents the means and standard deviations of these slopes. For third graders, the average slopes were not statistically significantly different from zero. For fourth graders, all but the slope for Specificity were lower than zero ($ps < .01$), indicating that students declined in both their calibration and their accuracy across the year.

{Insert Table 4.13}

Results from regressions of growth of posttest accuracy on growth of pretest calibration are presented in Table 4.14. No statistically significant associations emerged between growth in calibration and growth in accuracy when calculated as slopes from regressions of each on time (represented as quiz number). Neither average pretest accuracy nor assignment to ETG predicted this measure of performance growth. In the expanded table within the Appendix (Table 15), only three variables showed any statistically significant association with quiz performance growth: students in fourth grade, Asian students, and those eligible for free/reduced lunch showed lower levels of growth in posttest performance. Regressions of end-of-year math CST scores on calibration slopes painted a similar picture with regard to all but one of the calibration slopes (Table 4.15). Growth in Simple Match was associated with higher math test scores ($\beta = .03$, $p = .017$). After using a Bonferroni correction (Abdi, 2007) to adjust the necessary p value for the five comparisons conducted in this set of analyses, the required p value of .01 was not obtained. Other variables from these regressions did display statistically significant associations with math

CST scores: average quiz pretest accuracy, whether the student was a male, and prior math and ELA CST scores were all positively associated with performance (Appendix C, Table 16).

{Insert Table 4.14}

{Insert Table 4.15}

Discussion

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration? As the context for practice and feedback, this study used ST Math, a digital mathematics learning environment focused on teaching math concepts through spatial representations. Within ST Math, students took quizzes on the content as they progressed through a number of mathematics objectives. For each quiz question, students gave their answer and rated their confidence; at the end of each quiz they were shown graphical feedback on the agreement between their accuracy and confidence. These training elements provided direct, individualized feedback targeting a specific aspect of metacognition: the accuracy of metacognitive judgments, often termed *calibration*. It was hypothesized that after a year's training on calibration within ST Math, students would display greater calibration accuracy than their peers who were randomly assigned to begin ST Math the following year. Analyses comparing those who started ST Math first (the ETG) to those with a delayed start (the LTG) did not support this hypothesis. Across multiple measures, the ETG displayed lower levels of calibration than the LTG.

A number of robustness checks investigated the stability of these differences. As different levels of mathematics knowledge were likely to influence levels of calibration (see Efklides, 2008; Study 2), measures of this prior knowledge were important to the analyses. Concerns over the endogeneity of previous year math achievement measures for the ETG were assuaged when

coefficients changed little with the inclusion of a measure prior to treatment for both groups. Additionally, different operationalizations of quiz calibration balanced novelty of the treatment for the LTG and the stability offered by the use of a greater number of contributing quizzes. Estimates changed little across operationalizations, and with regards to most measures of calibration, painted a consistent picture: The ETG had lower levels of four of the five measures of calibration, Sensitivity, Simple Match, Gamma, and Discrimination.

The measures of Simple Match, Gamma, and Discrimination are based on the assumption of a single process for metacognitive judgments (see Schraw et al., 2013). Within this view of metacognition, monitoring judgments about correct and incorrect performance are part of the same process (Nelson & Narens, 1990; Schraw et al., 2013). This is reflected in the formulas for calculating these measures: they combine accurate confidence judgments for both correct and incorrect measures. In contrast, Sensitivity and Specificity assume separate processes for correct versus incorrect performance and divide these with formulas that represent the process of making correct judgments of confidence in Sensitivity and correct judgments of uncertainty in Specificity (Schraw et al., 2013). In this study, the ETG had lower calibration than the LTG when calibration was represented either as a combined measure or as Sensitivity. With respect to Specificity, however, the finding was opposite: the ETG made more judgments of uncertainty when they were incorrect than did the LTG. This difference suggests that there are separate calibration processes for correct and incorrect judgments (Feuerman & Miller, 2008; Schraw et al., 2013), and that the results for the three combined measures may present diluted versions of the Sensitivity finding. This is supported by the consistently lower beta coefficients for the combined measures than those for Sensitivity.

Focusing on Sensitivity and Specificity presents a clearer and more fine-grained picture of student calibration in this study. After a year's practice and feedback with calibration, the ETG was less likely to indicate confidence on items they correctly answered and more likely to indicate uncertainty on items they incorrectly answered. What aspects of the program may have resulted in this finding? Prior research supports the idea that students who are new to a domain are often overconfident, underestimating the demands of the task and failing to identify important aspects of the problem (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Kruger & Dunning, 1999). Those students who had been exposed to the format of instruction and testing within ST Math and who had been given feedback on their accuracy and calibration may have better recognized the complexity of the problems and displayed greater doubt. It is unclear whether this calibration pattern is adaptive, especially because it appeared they displayed this doubt for both correct *and* incorrect solutions. Findings from Study 2 suggest that confidence for correct answers (Sensitivity) is an important aspect of calibration and plays a larger role than Specificity in student learning. Within Study 2, student-specific characteristics were controlled as part of the analysis, but the effect of calibration may have been biased by characteristics of the Task X Person level (Efklides, 2008, 2011), including student overall confidence for the particular content within a given objective in ST Math. It may be that positive (and even inflated) confidence is important for a broader level of content (e.g., math generally, ST Math generally, fractions generally), but less important at the individual item level, where accurate Specificity may more readily engage control processes (see Bandura, 1986; Efklides 2008, 2011). The specific division for this change in the importance of confidence is unknown; also unknown is the level at which ST Math is operating to influence changes in calibration. Prior work indicates that ST Math has positive effects on general mathematics self-efficacy (Chang, Rutherford, &

Farkas, 2014; Rutherford, Hinga, Chang, Conley, & Martinez, 2011)—this indicates that the decrease in confidence and increase in uncertainty is at finer level than mathematics in general.

ST Math in the context of other training programs. Prior studies specifically focusing on the improvement of calibration have largely been situated in college classes such as Educational Psychology (e.g., Bol & Hacker, 2001; Schraw, Potenza, & Nebelsick-Gullet, 1993). Practice with multiple related quizzes within highly-related content area (herein, mathematics) was theorized to induce stronger transfer of metacognitive skills between quizzes than typically seen in classes with more unrelated content across exams. The conceptualization of elementary mathematics content as highly related, at least within the ST Math curriculum, may have been unfounded—the correlation between objective quizzes was low, with most correlations below .2.

The length of the training was also theorized to strengthen effects. Even though the content may not have been as related as expected, students did get a substantial amount of practice with quizzes and confidence judgments. On average, students completed more than 14 objectives and the corresponding pre/posttest quiz pairs. Although just over half of the available 23 quizzes, this number of quizzes is far greater than those used in previous studies (e.g., Bol et al., 2005; Huff & Nietfeld, 2009). Although the length of the training likely provided sufficient practice time as recommended by Bannert and Mengelkamp (2013), the training did not meet one of their other guidelines for effective metacognitive support programs. ST Math did not provide explicit instruction on calibration. It may be, especially with the young children in this study, that more explanation was needed to get them to see the value and understand the use of the confidence judgments they were making. Students in ST Math were given the opportunity to graphically view their calibration results (see Figure 4.1), but this may not have been enough. The program may have needed to direct the students and monitor their engagement with the

calibration feedback as in Labuhn and colleagues (2010) where students worked one-on-one with a researcher to graph their calibration over time. Although ST Math provided more opportunity to view and engage with feedback than in Labuhn et al., students did not have to click on their calibration results nor did they have to monitor their improvements in calibration. MIND Research Institute is currently investigating changes to the ST Math interface that will allow the tracking of these student click behaviors and will present more engaging information on calibration trends for the students. These changes will take the software closer to the treatment provided in Labuhn et al. (2010), and as a result, may show greater effectiveness.

Limitations to conclusions. Considering this study as a post-test only design with the assumption of successful randomization, results indicated that ST Math had a negative effect on student Sensitivity (confidence when correct) and a positive effect on student Specificity (uncertainty when incorrect). However, examination of the ETG's calibration during their first year of treatment indicated that the two groups may have been different before the variation in training was introduced. The ETG had lower levels of Sensitivity and all combined measures of calibration at the beginning of their first year and throughout their first year than did the LTG during their first year. Differences in combined measures of calibration diminished between years one and two, whereas differences in Sensitivity and Specificity increased: the ETG had lower levels of Sensitivity in their second year than their first and higher levels of Specificity. This supports the conclusion that the ETG was becoming more uncertain as they interacted with the software, but because there is no true pretest of calibration taken at the same time and during the same grade levels for both groups, it is difficult to say whether differences were due to exposure to ST Math or to other student or class characteristics.

(2) Is improvement in calibration accuracy linked to improvement in performance?

Few prior studies showing improvement in calibration have shown corresponding gains in achievement (*cf* Nietfeld et al., 2006). Some of the same features noted as contributing to the potential for calibration training within ST Math were also thought to increase the potential for calibration improvements to translate to achievement gains. These features were the length of training and hypothesized close relation between math content objectives. As noted above, the objectives were not as closely related as expected. Because of this, the first analysis for question 2 focused on pairs of more related quizzes: one early in the curriculum and one late in the curriculum. Even with these related quizzes, for the combined measures of calibration, growth of calibration was not related to growth in performance. The disaggregated measures, Sensitivity and Specificity, did show statistically significant associations with performance gain: student gain in Sensitivity (confidence when correct) was associated with positive gains in posttest accuracy, whereas gain in Specificity (uncertainty when incorrect) was associated with decreases in posttest accuracy. Using all quiz data in the models with calculated slopes for calibration and performance gains did not replicate this finding: all calibration gain slopes were associated with positive, but not statistically significant, increases in performance slopes.

Neither set of analyses indicated that there were relations between increases in calibration and improvements in CST performance. Only one relation between slope of calibration improvement and CST improvement emerged as statistically significant, that for Simple Match. However, when p levels were adjusted for multiple comparisons, the new threshold was not met. One possibility for this lack of association may be a disconnect between content within the CST and content within ST Math. It appears though that CSTs and ST Math *were* related: slopes of ST Math pretest accuracy improvement had moderate associations with CST improvement

($\beta=.21$, from Table 4.15), indicating that those students who improved in their accuracy across ST Math in 2011 also improved their math CST scores from the end of 2010 to the end of 2011.

Limitations to conclusions. Because of the study design and the nature of calibration measures, it was not possible directly test the hypothesis that improvements in calibration due to the treatment led to improvement in achievement. However, an attempt was made to move beyond the typical single-timepoint correlational research within calibration studies and explore the associations between change in calibration and change in achievement, regardless of treatment group. This provided interesting information on these associations, but could not causally attribute change in achievement to change in calibration (or rule out this relation). In order to directly test the mediation hypothesis, pretest calibration measures were needed for the LTG, either within or outside of ST Math. The current study took advantage of variation in treatment as part of a larger study on ST Math. As such, design decisions were not made to optimize ability to draw conclusions regarding calibration; future studies specifically designed for calibration research can correct this flaw.

Conclusion. It was thought that the lengthy and repeated practice on quizzes of related topics within ST Math would produce stronger effects on calibration than those typically seen in calibration research. This was not the case. Students who had a year with ST Math had lower levels of most measures of calibration than did students who were just starting ST Math, though ST Math did appear to have positive effects on students' indications of uncertainty for those questions they got wrong. It is unclear whether ST Math increased this specific kind of calibration, increased student uncertainty in general, or if the findings were a result of the different levels of calibration between the two groups before the intervention.

Students varied in their levels of calibration improvement across all five measures; however, this variation in gain was not associated with variation in achievement. There was no evidence to support the hypothesis that students who improved their calibration in ST Math would similarly improve their performance within ST Math or on an outside standardized achievement test. These findings shed light on both the potential processes of calibration and the difficulty in improving calibration.

More detailed work with measures of Sensitivity and Specificity at different task levels can extend the research from this study to separate the processes of certainty and uncertainty and further understand both their malleability and their relation with achievement across grain sizes. This study's failure to find calibration improvement or the link between calibration and achievement change can inform the design of both future interventions and studies. Future interventions with similar populations should increase the relation between content in quizzes and include explicit instruction directing students toward calibration monitoring and use. Future studies should measure calibration within and outside the intervention arena and take care to ensure equality of starting calibration between treatment groups. Calibration remains an important element of models of metacognition and SRL; careful study of how and when calibration works and can be changed can aid in the design of more successful interventions to improve calibration, contributing to greater student success in SRL, mathematics, and beyond.

References

- Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In Salkind, N. J. (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.
- Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3.
doi:10.1016/j.learninstruc.2012.10.003
- Azevedo, R. (2005). Using Hypermedia as a Metacognitive Tool for Enhancing Student Learning? The Role of Self-Regulated Learning. *Educational Psychologist, 40*(4), 199–209. doi:10.1207/s15326985ep4004_2
- Azevedo, R. (2007). Understanding the complex nature of self-regulatory processes in learning with computer-based learning environments: an introduction. *Metacognition and Learning, 2*(2-3), 57–65. doi:10.1007/s11409-007-9018-5
- Azevedo, R., & Bernard, R. M. (1995). A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research, 13*(2), 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT
- Azevedo, R., & Hadwin, A. F. (2005). Scaffolding Self-regulated Learning and Metacognition – Implications for the Design of Computer-based Scaffolds. *Instructional Science, 33*(5-6), 367–379. doi:10.1007/s11251-005-1272-9
- Bandura, Albert. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs N.J.: Prentice-Hall.
- Bannert, M., & Mengelkamp, C. (2013). Scaffolding Hypermedia Learning Through Metacognitive Prompts. In R. Azevedo & V. Aleven (Eds.), *International Handbook of*

- Metacognition and Learning Technologies* (pp. 171–186). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-5546-3_12
- 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(2), 133–151. doi:10.2307/20152656
- 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(4), 269–290. doi:10.2307/20157403
- California Department of Education (2011). California English Language Development Test [Test website]. Retrieved from <http://www.cde.ca.gov/ta/tg/el/>
- California State Board of Education (2010a). *K-12 California's common core content standards for mathematics*. Retrieved on March 19, 2011 from http://www.scoe.net/castandards/agenda/2010/math_ccs_recommendations.pdf
- California State Board of Education (2010b). *Technical Q & A: Percent proficient*. Retrieved from <http://www.cde.ca.gov/ta/ac/ap/techqa06b.asp>
- Chang, A., Rutherford, T., & Farkas, G. (2014, April). *I can do it!: Expectancy as a mediator of the ST Math effect on math achievement*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: the learning trajectories approach*. Taylor & Francis.
- Crocker, J., Voelk, K, Testa, M. & Major, B. (1991) Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology*, 60(1), 218-228.

- Dabbagh, N., & Kitsantas, A. (2005). Using Web-based Pedagogical Tools as Scaffolds for Self-regulated Learning. *Instructional Science*, 33(5-6), 513–540. doi:10.1007/s11251-005-1278-3
- De Corte, E., Verschaffel, L., & Op ’t Eynde, P. (2005). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation*. Elsevier.
- Dembo, M. H., & Jakubowski, T. G. (2003). *The influence of self-protective perceptions on the accuracy of test prediction*. Presented at The Annual conference of the American Educational Research Association, Chicago, IL.
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(3), 231–264. doi:10.1007/s11409-008-9029-x
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students’ judgments can improve their achievement. *Learning and Instruction*, 24, 58–61. doi:10.1016/j.learninstruc.2012.05.002
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science*, 12(3), 83–87. doi:10.1111/1467-8721.01235
- Educational Testing Service (2008). *California Standards Tests (CSTs) technical report, spring 2007 administration*. Retrieved from <http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt07.pdf>

- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287.
doi:10.1027/1016-9040.13.4.277
- Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist, 46*(1), 6–25.
doi:10.1080/00461520.2011.538645
- Greene, J. A., & Azevedo, R. (2007). A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions. *Review of Educational Research, 77*(3), 334–372. doi:10.3102/003465430303953
- Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 429-455). New York: Psychology Press
- Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170.
doi:http://dx.doi.org/10.1037/0022-0663.92.1.160
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487
- Hoyt, C. L., Aguilar, L., Kaiser, C. R., Blascovich, J., & Lee, K. (2007). The self-protective and undermining effects of attributional ambiguity. *Journal of Experimental Social Psychology, 43*(6), 884–893. doi:10.1016/j.jesp.2006.10.013
- Huff, J., & Nietfield, J. (2009). Using Strategy Instruction and Confidence Judgments to Improve Metacognitive Monitoring. *Metacognition and Learning, 4*(2), 161–176.

- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3), 217–273. doi:10.1016/0001-6918(91)90036-Y
- Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an On-Line Formative Assessment and Tutoring System. *Journal of Educational Computing Research*, 43(4), 489–510.
- Kramarski, B., & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning*, 22(1), 24–33. doi:10.1111/j.1365-2729.2006.00157.x
- Kramarski, B., & Zeichner, O. (2001). Using Technology to Enhance Mathematical Reasoning: Effects of Feedback and Self-Regulation Learning. *Educational Media International*, 38(2-3), 77–82. doi:10.1080/09523980110041458
- 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(6), 1121–1134.
- Labuhn, A., Zimmerman, B., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition and Learning*, 5(2), 173–194. doi:10.1007/s11409-010-9056-2
- Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A Comparison of Traditional Homework to Computer-Supported Homework. *Journal of Research on Technology in Education*, 41(3), 331–359.
- Nelson, T. O. & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. In Gordon H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 26, pp.

- 125–173). Academic Press. Retrieved from
<http://www.sciencedirect.com/science/article/pii/S0079742108600535>
- 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*(2), 159–179. doi:10.1007/s10409-006-9595-6
- Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40. doi:10.1037/0022-0663.82.1.33
- Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... Gallagher, L. P. (2010). Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics. *American Educational Research Journal, 47*(4), 833–878. doi:10.3102/0002831210367426
- Rutherford, T., Hinga, B., Chang, A., Conley, A. M., & Martinez, M. E. (2011, August). *The effect of ST Math software on standardized test scores via improvement in mathematics expectancy*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.
- Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Graham, J., Kibrick, M., ... Martinez, M. E. (2014) A randomized trial of an elementary school mathematics software intervention: Spatial-Temporal (ST) Math. *Journal of Research on Educational Effectiveness*. doi: 10.1080/19345747.2013.856978
- Schenke, K., Rutherford, T., & Farkas, G. (2014) Alignment of game design features and state mathematics standards: Do results reflect intentions? *Computers & Education*. doi: 10.1016/j.compedu.2014.03.019

- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57.
doi:10.1016/j.learninstruc.2012.08.007
- Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the Calibration of Performance. *Contemporary Educational Psychology, 18*(4), 455–463.
doi:10.1006/ceps.1993.1034
- Stone, N. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. doi:10.1023/A:1009084430926
- Winne, P. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466–488.
- Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational Psychology Review, 20*(4), 429–444. doi:10.1007/s10648-008-9080-9
- Zimmerman, B. J., & Kitsantas, A. (2014). Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. *Contemporary Educational Psychology, 39*(2), 145–155. doi:10.1016/j.cedpsych.2014.03.004
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. doi:10.3102/0002831207312909

The screenshot displays a quiz interface with two main panels. The left panel contains a question: "1 What is the value of 4 in 40,892?". Below the question are four multiple-choice options: (A) 400, (B) 4,000, (C) 40,000, and (D) 400,000. Option (C) is selected. A "Confidence Level" section at the bottom of this panel shows "I'm not sure" and two icons: a shrugging person and a cheering person. The right panel, titled "Quiz Results", shows a score of $\frac{2}{5}$ (40%) and states "2 out of 5 questions were correct. 3 questions were incorrect." Below this is a list of five questions with their answers and correctness: 1. C ✓, 2. C ✗, 3. B ✓, 4. C ✗, 5. C ✗. A "High Confidence Score" box shows $\frac{2}{4} = 50\%$. At the bottom right of the results panel are icons for a plus sign and a cheering person. The text "Using Place Value" is visible at the bottom of the right panel.

Figure 4.1. Screen shot from an ST Math quiz. On the left, the student chooses his/her answer and then indicates how sure they are of the answer by selecting the shrugging icon or the cheering icon. On the right, the student sees of the ones he/she felt sure (confident), how many he/she actually got correct.

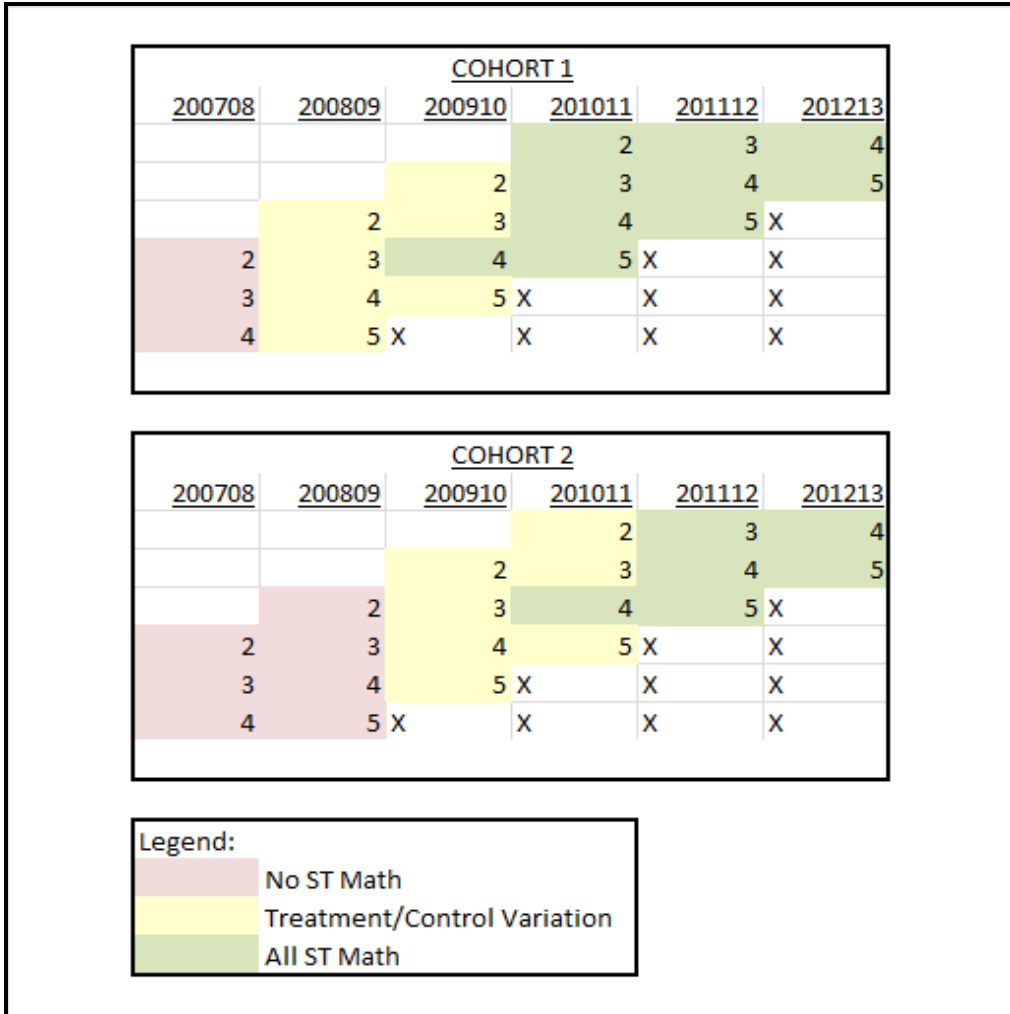


Figure 4.2. ST Math study design. Illustrates the progression of students within a given grade as the years continue. "X" marks indicate the cohort of students has aged out of the study.

Table 4.1

Sample Demographics, Divided by Treatment Group

	ETG	LTG
Grade 3	51%	47%
Grade 4	49%	53%
Male	52%	52%
Asian	4%	3%
Hispanic	64%	80%
White	27%	15%
Other Ethnicity	4%	2%
English Language Learner	60%	63%
Free/Reduced Lunch	81%	86%
N	1,272	1,353

Table 4.2a

Descriptive Statistics of Calibration and Achievement Variables, Third Grade by Treatment Group

	ETG		LTG		Difference	p value
	Mean	SD	Mean	SD		
Total Objectives	15.81	6.53	14.71	6.90	1.10	0.004
Pretest Accuracy	0.62	0.14	0.67	0.15	-0.05	<.0001
Posttest Accuracy	0.74	0.13	0.78	0.13	-0.04	<.0001
Pretest Sensitivity	0.79	0.21	0.86	0.16	-0.07	<.0001
Posttest Sensitivity	0.85	0.20	0.90	0.15	-0.05	<.0001
Pretest Specificity	0.38	0.23	0.35	0.21	0.03	0.005
Posttest Specificity	0.38	0.20	0.38	0.16	0.00	0.611
Pretest Simple Match	0.66	0.12	0.70	0.13	-0.04	<.0001
Posttest Simple Match	0.73	0.13	0.78	0.12	-0.05	<.0001
Pretest Gamma	0.36	0.27	0.43	0.28	-0.07	<.0001
Posttest Gamma	0.49	0.29	0.59	0.24	-0.10	<.0001
Pretest Discrimination	0.97	0.77	1.14	0.80	-0.17	0.0001
Posttest Discrimination	1.29	0.79	1.57	0.72	-0.28	<.0001
ELA CST Score 2011	348.47	60.03	356.79	55.67	-8.32	0.010
Math CST Score 2011	383.14	76.40	382.00	77.68	1.14	0.792
ELA CST Score 2012	338.68	59.57	333.16	54.67	5.52	0.010
Math CST Score 2012	398.63	79.05	394.93	76.75	3.70	0.792
N	649		630			

Note. ETG is Early Treatment Group, LTG is Late Treatment Group. P values calculated from t-test of differences of means, assuming unpaired data. Positive differences indicate higher values for the ETG.

Table 4.2b

Descriptive Statistics of Calibration and Achievement Variables, Fourth Grade by Treatment Group

	ETG		LTG		Difference	p value
	Mean	SD	Mean	SD		
Total Objectives	14.59	5.99	14.50	6.40	0.09	0.791
Pretest Accuracy	0.52	0.15	0.54	0.14	-0.02	0.055
Posttest Accuracy	0.66	0.15	0.67	0.14	-0.01	0.139
Pretest Sensitivity	0.73	0.22	0.79	0.18	-0.06	<.0001
Posttest Sensitivity	0.81	0.2	0.86	0.16	-0.05	<.0001
Pretest Specificity	0.41	0.25	0.35	0.22	0.06	<.0001
Posttest Specificity	0.39	0.21	0.35	0.19	0.04	0.0004
Pretest Simple Match	0.61	0.13	0.61	0.12	0.00	0.63
Posttest Simple Match	0.68	0.14	0.70	0.12	-0.02	0.016
Pretest Gamma	0.27	0.28	0.26	0.28	0.01	0.99
Posttest Gamma	0.41	0.29	0.44	0.27	-0.03	0.015
Pretest Discrimination	0.73	0.74	0.72	0.76	0.01	0.917
Posttest Discrimination	1.10	0.79	1.17	0.75	-0.07	0.075
ELA CST Score 2010 ^a	342.10	60.59	345.72	56.88	-3.62	0.279
Math CST Score 2010 ^a	362.83	73.38	365.01	77.30	-2.18	0.613
ELA CST Score 2011	322.45	58.89	331.43	55.62	-8.98	0.004
Math CST Score 2011	385.56	83.88	389.06	85.02	-3.50	0.449
ELA CST Score 2012	354.87	54.51	359.85	54.63	-4.98	0.004
Math CST Score 2012	387.02	69.94	392.02	73.45	-5.00	0.449
	N	623		723		

Note. ETG is Early Treatment Group, LTG is Late Treatment Group. P values calculated from t-test of differences of means, assuming unpaired data. Positive differences indicate higher values for the ETG.

^aNot all students had data on this variable, N=571 for the ETG and N=667 for the LTG.

Table 4.3

Correlations between Calibration Measures and Achievement Measures

N=2,625	Pretest Calibration Measures								
	Pretest Accuracy	Posttest Accuracy	Sensitivity	Specificity	Match	Gamma	Discrim.	Math CST	ELA CST
Pretest Accuracy	1	0.783	0.338	0.024 ^a	0.731	0.610	0.543	0.461	0.559
Posttest Accuracy	0.783	1	0.282	0.014 ^a	0.234	0.239	0.192	0.519	0.586
Sensitivity	0.284	0.335	0.807	-0.765	0.234	0.239	0.192	0.124	0.163
Specificity	0.059 ^c	0.054 ^c	-0.699	0.794	0.314	0.358	0.475	0.016 ^a	0.050 ^b
Simple Match	0.614	0.730	0.537	0.089	0.682	0.905	0.852	0.358	0.462
Gamma	0.530	0.617	0.489	0.188	0.918	0.522	0.923	0.269	0.367
Discrimination	0.486	0.555	0.406	0.356	0.858	0.923	0.493	0.222	0.327
Math CST	0.461	0.519	0.186	0.014 ^a	0.399	0.331	0.290	1	0.721
ELA CST	0.559	0.586	0.206	0.059 ^c	0.477	0.399	0.374	0.721	1

Note. All correlations at $p < .001$ level except ^a $p > .05$, ^b $p < .05$, ^c $p < .01$. Correlation between pretest measures and between pretest measures and achievement above the diagonal. Correlation between posttest measures and between posttest measures and achievement below the diagonal. Correlation between same measure pre to posttest on the diagonal (shaded). Data from aggregations across all quizzes taken in 2011. N for correlations with ELA CST 2,624 (1 missing student).

Table 4.4

Effect of Early Treatment Group on Calibration for Place Value

		(1)	(2)	(3)	(4)	(5)
N=2,560		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
ETG	B	-0.05***	0.01	-0.05***	-0.09***	-0.24***
	SE	(0.01)	(0.01)	(0.01)	(0.02)	(0.06)
	Beta	-0.12***	0.02	-0.10***	-0.08***	-0.07***
Pretest Acc	B	0.12***	-0.01	0.17***	0.28***	0.70***
	SE	(0.02)	(0.03)	(0.02)	(0.04)	(0.12)
	Beta	0.14***	-0.01	0.19***	0.14***	0.11***
Constant	B	0.75***	0.20**	0.37***	-0.13	-0.48
	SE	(0.05)	(0.05)	(0.04)	(0.08)	(0.23)
R2		0.07	0.02	0.19	0.11	0.08

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included math and ELA CST scores, objective pretest quiz accuracy, grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). Sample limited to those who had data on the Place Value objective, the first objective for all third graders and first or second objective for all fourth graders.

Table 4.5

Effect of Early Treatment Group on Calibration for First Three Objectives Encountered in ST Math

		(1)	(2)	(3)	(4)	(5)
N=2,624		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
ETG	B	-0.06***	0.02	-0.04***	-0.08***	-0.22***
	SE	(0.01)	(0.01)	(0.01)	(0.02)	(0.05)
	Beta	-0.14***	0.04	-0.13***	-0.10***	-0.10***
Pretest Acc	B	0.19***	0.03	0.28***	0.54***	1.34***
	SE	(0.03)	(0.04)	(0.03)	(0.05)	(0.17)
	Beta	0.19***	0.03	0.33***	0.28***	0.24***
Constant	B	0.58***	0.19	0.10	-0.58**	-1.40**
	SE	(0.12)	(0.10)	(0.10)	(0.19)	(0.44)
R2		0.13	0.05	0.32	0.20	0.16

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included math and ELA CST scores, objective pretest quiz accuracy, grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). Sample limited to those who had data on at least three objectives, the specific objectives included were controlled as a series of dummy variables (omitted). One student who otherwise had the appropriate data was excluded from this analysis for missing ELA scores.

Table 4.6

Effect of Early Treatment Group on Calibration Aggregated Across Entire Year

N=2,624		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
ETG	B	-0.04**	0.02*	-0.02**	-0.04**	-0.10**
	SE	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
	Beta	-0.11**	0.06*	-0.08**	-0.07**	-0.06**
Pretest Acc	B	0.28***	0.05	0.41***	0.79***	1.99***
	SE	(0.03)	(0.04)	(0.03)	(0.06)	(0.17)
	Beta	0.24***	0.04	0.48***	0.43***	0.39***
Constant	B	0.65***	0.32***	0.35***	-0.18***	-0.28
	SE	(0.04)	(0.04)	(0.03)	(0.04)	(0.14)
R2		0.14	0.06	0.43	0.31	0.27

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included math and ELA CST scores, objective pretest quiz accuracy, grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). The specific objectives included were controlled as a series of dummy variables (omitted). One student who otherwise had the appropriate data was excluded from this analysis for missing ELA scores.

Table 4.7

Descriptive Statistics for Calibration and Accuracy Across Years, Early Treatment Group

	2010		2011		2010		2011	
	Grade 2		Grade 3		Grade 3		Grade 4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest Acc. ^{ab}	0.59	0.14	0.63	0.14	0.59	0.16	0.52	0.15
Sensitivity ^b	0.83	0.18	0.85	0.20	0.85	0.18	0.81	0.20
Specificity ^{ab}	0.36	0.21	0.38	0.20	0.34	0.20	0.39	0.21
Simple Match ^a	0.69	0.13	0.73	0.13	0.69	0.14	0.68	0.14
Gamma ^a	0.42	0.27	0.49	0.29	0.41	0.29	0.41	0.29
Discrimination ^a	1.11	0.77	1.30	0.79	1.11	0.80	1.10	0.79
Posttest Acc. ^a	0.68	0.14	0.74	0.13	0.67	0.16	0.66	0.15
N	644		643		615		616	

Note. ^aDifferences between years for those starting in second grade ($p < .05$). ^bDifferences between years for those starting in third grade ($p < .05$). Limited to those students present in both 2010 and 2011 data (excluding 1%). From grade-level Ns, one student skipped third grade in 2011.

Table 4.8
*Standardized Regression Coefficients Compared Across Analyses and Samples:
 Association of Treatment Group and Measures of Calibration*

	(1) Sensitivity	(2) Specificity	(3) Simple Match	(4) Gamma	(5) Discrimination
<i>Total Sample</i>					
Year 1 vs 1	-0.10*	0.02	-0.13***	-0.11***	-0.10**
Year 2 vs 1	-0.11**	0.06*	-0.08**	-0.07**	-0.06**
<i>Limited to Only 2011 Fourth Graders, with Pretest CST Scores</i>					
Year 1 vs 1	-0.08	0.04	-0.08	-0.07	-0.04
Year 2 vs 1	-0.13***	0.10**	-0.05	-0.04	-0.03

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Standardized regression coefficients taken from Appendix C, Table 10 (top line), Table 5 (second line), Appendix C, Table 11 (third line), and Appendix C, Table 7 (fourth line).

Table 4.9
Sample Characteristic Comparison for Question 2

	Excluded	Included
Grade 3	58%	43%
Grade 4	42%	57%
Male	47%	55%
Hispanic	76%	71%
Asian	1%	5%
White	21%	21%
Other Ethnicity	2%	4%
English Language Learner	74%	54%
Free/Reduced Lunch	88%	81%
N	959	1586

Table 4.10

Descriptive Statistics of Selected Starting and Ending Quiz Pairs

	3rd Grade			4th Grade		
	Start	End	Gain	Start	End	Gain
Pretest Acc.	0.67 (0.28)	0.83 (0.22)	0.16 (0.29)	0.62 (0.22)	0.47 (0.23)	-0.15 (0.25)
Posttest Acc.	0.75 (0.25)	0.82 (0.19)	0.07 (0.26)	0.70 (0.21)	0.60 (0.23)	-0.10 (0.24)
Sensitivity	0.82 (0.30)	0.90 (0.23)	0.08 (0.31)	0.87 (0.25)	0.71 (0.37)	-0.16 (0.38)
Specificity	0.36 (0.37)	0.36 (0.30)	0.00 (0.44)	0.34 (0.37)	0.39 (0.40)	0.05 (0.44)
Simple Match	0.69 (0.26)	0.79 (0.24)	0.10 (0.30)	0.67 (0.21)	0.58 (0.22)	-0.09 (0.25)
Gamma	0.40 (0.66)	0.59 (0.55)	0.19 (0.77)	0.46 (0.56)	0.20 (0.63)	-0.26 (0.76)
Discrimination	1.05 (1.90)	1.46 (1.63)	0.41 (2.36)	1.11 (1.59)	0.50 (1.46)	-0.61 (1.99)
Math CST	411.22 (65.75)	436.73 (68.55)	25.51 (60.59)	411.78 (79.42)	413.23 (64.93)	1.45 (64.46)
ELA CST	372.74 (52.40)	357.81 (53.21)	-14.93 (37.02)	339.43 (55.04)	369.61 (51.02)	30.18 (35.48)
N	683			903		

Note. Means from starting and ending pairs of selected objective quizzes and from CST scores in 2011 (start) and 2012 (end). Standard deviations in parentheses. Gains and standard deviations of gains provided.

Table 4.11

Association between Calibration Gain and Posttest Performance Gain, Paired Quizzes

N=1,586		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Gain	B	0.05*	-0.04**	-0.03	0.0001	-0.001
	SE	(0.02)	(0.01)	(0.03)	(0.01)	(0.003)
	Beta	0.07*	-0.07**	-0.04	0.0001	-0.005
Pretest Gain	B	0.22***	0.24***	0.25***	0.23***	0.23***
	SE	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
	Beta	0.22***	0.24***	0.25***	0.23***	0.23***
Early Treat Group	B	0.03*	0.03*	0.04**	0.03**	0.03**
	SE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Beta	0.03*	0.03*	0.04**	0.03**	0.03**
Constant	B	0.01	0.01	0.01	0.01	0.01
	SE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
R2		0.17	0.17	0.17	0.16	0.16

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables not shown included grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18).

Table 4.12

Association between Calibration Gain and Math CST Gain

		(1)	(2)	(3)	(4)	(5)
N=1,586		Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration	B	-8.19	5.64	2.53	-2.31	-0.17
Gain	SE	(5.55)	(4.21)	(4.33)	(2.39)	(0.64)
	Beta	-0.05	0.04	0.01	-0.03	-0.01
Early Treat	B	-1.16	-1.26	-1.74	-1.48	-1.60
Group	SE	(5.68)	(5.69)	(5.57)	(5.65)	(5.66)
	Beta	-0.01	-0.01	-0.01	-0.01	-0.01
Constant	B	17.60	17.05	17.08	17.68	17.35
	SE	(8.92)	(9.10)	(9.12)	(8.98)	(9.08)
	R2	0.05	0.06	0.05	0.05	0.05

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables not shown included grade, gender, race, language and free/reduced priced lunch statuses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18).

Table 4.13

Slopes for Improvement in Accuracy and Calibration over Time

N=2,625	3rd Grade		4th Grade	
	Mean(B)	SD	Mean(B)	SD
Posttest Accuracy	0.002	0.042	-0.011***	0.046
Sensitivity	0.001	0.066	-0.007***	0.052
Specificity	-0.002	0.057	0.002	0.057
Simple Match	0.0004	0.038	-0.005***	0.041
Gamma	-0.001	0.112	-0.011**	0.110
Discrimination	-0.006	0.294	-0.030***	0.284
N	1,279		1,346	

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. P values indicate whether statistically significantly different from zero.

Table 4.14

Association between Calibration Growth and Growth of Quiz Posttest Performance

		(1)	(2)	(3)	(4)	(5)
N=2,625		Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Slope	B	0.04	0.05	0.18	0.06	0.02
	SE	(0.12)	(0.12)	(0.12)	(0.04)	(0.02)
	Beta	0.05	0.06	0.16	0.15	0.15
Pretest Acc.	B	-0.02	-0.02	-0.02	-0.02	-0.02
	SE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Beta	-0.06	-0.06	-0.06	-0.06	-0.06
Early Treat. Group	B	-0.001	-0.00	-0.001	-0.001	-0.001
	SE	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	Beta	-0.01	-0.01	-0.01	-0.02	-0.02
Constant	B	0.01	0.01	0.01	0.01	0.01
	SE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	R2	0.07	0.07	0.09	0.09	0.09

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables not shown included grade, gender, race, language and free/reduced priced lunch statuses. Specific objectives tested controlled with a series of dummy variables (not shown). The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18).

Table 4.15

Association between Calibration Growth and End-of-Year Math CST Scores

		(1)	(2)	(3)	(4)	(5)
N=2,624		Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Slope	B	-0.90	17.11	48.92*	9.75	3.06
	SE	(10.44)	(9.83)	(18.40)	(6.98)	(2.60)
	Beta	-0.001	0.01	0.03*	0.01	0.01
Pretest Acc.	B	100.49***	100.24***	100.22***	100.40***	100.39***
	SE	(10.97)	(10.92)	(10.75)	(10.92)	(10.87)
	Beta	0.21***	0.21***	0.21***	0.21***	0.21***
Early Treat. Group	B	1.86	1.87	1.82	1.78	1.80
	SE	(2.82)	(2.84)	(2.82)	(2.83)	(2.84)
	Beta	0.01	0.01	0.01	0.01	0.01
Constant	B	105.76***	105.61***	104.64***	105.18***	105.50***
	SE	(17.53)	(17.46)	(17.19)	(17.49)	(17.49)
	R2	0.64	0.64	0.64	0.64	0.64

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables not shown included grade, prior CST scores gender, race, language and free/reduced priced lunch statuses. Specific objectives included controlled with a series of dummy variables (not shown). The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). One student omitted who did not have ELA CST data for 2011.

CHAPTER FIVE

Summary and Conclusion

Self-Regulated Learning (SRL), the ability to set goals, monitor progress toward these goals, and make adjustments when necessary, is an important part of a positive mathematical disposition (DeCorte, Verschaffel, & Op'T Eynde, 2000; Zimmerman, 2008). Within SRL, accurate metacognitive monitoring is necessary to drive control processes needed for adjustments (Efklides, 2008; Winne, 2004). Students who display this accuracy are said to be *calibrated*, and the role of calibration in SRL and learning in general has been a topic of growing interest within Educational Psychology (Alexander, 2013). This interest has been accompanied by a number of studies of calibration and its relation with achievement, but unanswered questions remain about the nature of calibration: how it should be measured, its role as a dynamic aspect of metacognition, and how best to improve it. The studies within this dissertation used a rich source of data on student calibration and achievement in mathematics to approach these questions and present results on calibration as representative of a complex system of metacognition related to both the person and the task.

Summary of Findings

In Study 1, I examined which measures of calibration could accommodate real-world data of accuracy and confidence judgments. Prior work examining differences between measures focused on simulated data—my results showed that the distribution of such data is different than the distribution of real data, and these differences have implications for the calculability of calibration measures. Because of the distribution of data, even with large numbers of quiz questions (over 100), two of the most popular measures of calibration, Gamma and Discrimination, could not be calculated for a sizable portion of students. Kappa, Simple Match,

Sokal Reverse, Sensitivity and Specificity were less affected by missing data and were calculable for nearly all students. However, this would change when using tests with questions numbers typically used in research or practice (<20): many more measures would be incalculable. Excluding individuals within incalculable measures, as is sometimes done, is likely to bias the analysis sample as individual combinations of accuracy and confidence, and any resulting missing data, will be related to personal characteristics.

Also within this study, I examined which of ten commonly used measures of calibration had the greatest predictive validity for performance gains from pre to posttest. All measures of calibration were highly correlated, except for Sensitivity and Specificity, which were strongly negatively correlated only with each other. All measures also explained similar amounts of variance in performance, but overall these amounts were small and most betas were below .10. Although there were only limited differences between models, a combined model with Sensitivity and Specificity explained the most variance, and in such a model, both Sensitivity and Specificity were positively associated with performance gain. This confirmed the simulated data results from Schraw, Kuch, and Gutierrez (2013).

In Study 2, Sensitivity and Specificity were used to examine how, within the same student, variation in calibration was related to variation in performance gain from pre to posttest. I found that using zero-order correlations inflated the relation between calibration and performance, especially for Sensitivity (proportion confident when correct). Looking only at within-student variation, both Sensitivity and Specificity had positive associations with performance gain, but both beta coefficients were below .10, and the coefficient for Specificity was a third the size of the coefficient for Sensitivity. This suggested that it was more important for students to be confident when correct than to be uncertain when incorrect—a finding contrary

to assumptions about the trigger of control processes from accurately identifying material not yet learned. There appeared to be an association between Specificity and performance gain that operated at the between-student level, but this estimate may have been biased by omitted variables.

Study 2 also investigated whether calibration and the benefit from calibration depended on student grade level. Across two samples of students, neither level of calibration nor the association between calibration and performance varied systematically with grade level. There were statistically significant differences between the grades in calibration values and in regression coefficients, but no clear patterns emerged.

Study 3 characterized the practice and feedback on calibration within ST Math as a calibration intervention and asked whether a year's worth of practice with this intervention would improve calibration. Random variation in the timing of the intervention was used to compare an early treatment group (ETG) to a late treatment group (LTG). Results indicated that after one year of ST Math, the ETG had lower levels of calibration than the LTG on all measures but Specificity; the ETG had higher levels of Specificity than the LTG. Examination of the prior year's calibration for the ETG suggested that this difference may not have been due to the intervention, but was instead likely due to pre-existing levels of calibration in the ETG. However, when these prior levels were taken into consideration, it did appear that the ETG became more uncertain after their year with ST Math.

Study 3 also addressed the potential link between improvement in calibration and improvement in achievement. Looking first at change in calibration predicting change in achievement using two of the most related quizzes, I found that only changes in Sensitivity and Specificity were associated with change in quiz posttest performance between the objectives.

The association with Sensitivity change was positive and Specificity change negative, indicating that as students became more confident in their correct answers, they became more accurate and as students became more uncertain about their incorrect answers, they became less accurate. These findings were not replicated when change in calibration was operationalized as a slope from beginning of ST Math to end. In these models, none of the calibration change measures had associations with achievement change. Neither did changes in calibration have associations with changes in math achievement outside of ST Math: both methods of analyses returned null results regarding the link between calibration and math performance on the California Standards Tests.

Themes

Each study within this dissertation answered specific questions regarding the nature of calibration. The three, taken together, not only cover a common topic area and source of data, but also provide insight into the process and measurement of calibration beyond the individual research questions. These insights are discussed as two themes.

Calibration likely reflects a dual process. As suggested by Schraw and colleagues (2013) and work from clinical research (e.g., Feuerman & Miller, 2008), Sensitivity (proportion confident when correct) and Specificity (proportion uncertain when incorrect) are separable as measurements and likely represent separable constructs. The predictive validity results from Study 1 supported this conclusion—the model including these two measures explained the most variance and each had higher beta coefficients than other measures. Within Study 2, each had distinct relations with performance gains. These differences were present at the within-student level, and differences were also present between students, where only Specificity exhibited a contextual effect. In Study 3, practice with ST Math had opposite effects on the two measures, reducing Sensitivity and increasing Specificity. The effect of ST Math on combined measures of

calibration reflected the effect on Sensitivity but was diluted—viewing calibration as a single process by using measures such as Simple Match, Gamma, and Discrimination confounded the two opposite findings.

In the process of making metacognitive judgments, when students ask themselves, “did I get this right?” they may use cues such as the ease of reaching the answer or affective experiences to determine their level of certainty (Efklides, 2008; Flavell, 1979). Metacognitive accuracy may therefore depend on student understanding of and use of these cues (De Bruin & van Gog, 2012). Schraw et al. (2013) characterized Sensitivity and Specificity as representing distinct processes for correct and incorrect answers, but instead, they could represent distinct categories of cues: those for certainty and those for uncertainty. Student attention to and interpretation of cues such as those encountered when things come easily may be different from those encountered when things are difficult, such as feelings of frustration or confusion. These processes may also differ depending on the type of problem. Much of calibration research that considers such cues is conducted on knowledge recall questions (e.g., Bjork, Dunlosky, & Kornell, 2013; Kleitman & Stankov, 2001), where the feeling of recognition as a student conjures an answer may be very different from the process of determining an answer in multi-step problem solving such as that found in mathematics (see Jonsson & Allwood, 2003).

Calibration reflects features of the Task x Person level and the Person level.

Returning to the model articulated in Efklides (2008, 2011), the monitoring and control inherent in metacognition and in SRL involves both the more stable characteristics of the person and the interaction between these characteristics and the task at hand. Study 2 demonstrated differences in how calibration is related to performance at these different levels by looking both within students across tasks and between students. Prior research has noted that students often make top

down decisions about calibration, first determining if the subject is one that they are good at and then only slightly adjusting their decisions by task-specific information, such as the cues like feelings of knowing (see Zhao & Linderholm, 2008). Study 2 demonstrated that there is variance between tasks in these decisions and that this variance is related to the gains students make in performance. However, it is unclear whether this variance is due to student responsiveness to cues or to stable characteristics of the student at a finer grain than the individual generally or the individual in math. It could be that students draw on judgments of their competence with fractions or with simple addition or with problems involving the number “5” to make their decisions. These representations of the task will vary between students, as will the usefulness of each representation.

The results of Study 3 indicated that ST Math may have increased student uncertainty and decreased student confidence. However, prior research had demonstrated the positive effect of ST Math on student self-beliefs for mathematics (Chang, Rutherford, & Farkas, 2014; Rutherford, Hinga, Chang, Conley, & Martinez, 2011). These results may seem at odds, but considered in light of the results from Study 2, suggest a division in the process of confidence judgments by granularity of task.

Future Research

This dissertation provided novel information on the measurement and malleability of calibration and the relation between calibration and achievement; it also suggested new avenues of inquiry regarding calibration, metacognitive processes, and SRL. I plan to pursue these avenues both within and outside of the ST Math learning environment; I detail my most proximal research priorities below.

Within ST Math. ST Math provides an opportunity to collect rich data on student learning with a large and diverse sample of students. The MIND Research Institute (MIND) is a dedicated collaborator and has already taken steps to improve the calibration feedback within ST Math based on initial results of this dissertation. My future research priorities within ST Math focus on three main areas: structure and measurement of calibration, saliency of feedback to students, and ability to exercise control.

Structure and measurement of calibration. Focusing on the simplicity of measures thought necessary for elementary-aged children, current measures of calibration within ST Math are dichotomous, only asking children to express whether they are confident or uncertain. Recent research with very young children indicates that even preschoolers can correctly understand and use scales with three options (Ghetti, Hembacher, & Coughlin, 2013). Replicating the analyses within this dissertation with more fine-grained measures of calibration may lead to different conclusions and provide more information on the differences between the processes around confidence and uncertainty.

Additionally, embedding calibration measures within the game-play of ST Math (instead of just within the pre and posttest quizzes) may better illuminate how metacognitive judgments relate to control processes as students are engaging in learning. Students may also provide valuable clues about their metacognitive processes as they make their answer choices and determinations of confidence. Collecting click data on students that change their answers and how this relates to their calibration can indicate whether students who behaviorally express uncertainty (by waffling in their answer choice) recognize this uncertainty in their choice of confidence judgment.

Saliency of feedback. Students in ST Math were provided the opportunity to view the information on their calibration and accuracy after each quiz, but no data were collected on their actual viewing practices. MIND is planning to collect click data on whether the students review this information and which questions they revisit after reviewing this information. I plan to analyze these data to determine if students who actively engage with the calibration feedback make greater improvements in their calibration as they progress through the ST Math curriculum. Additionally, I have been collaborating with MIND to design modifications to the software that will encourage students to track their calibration results and to monitor their improvement in calibration. Once a prototype is designed, it will be pilot tested before rolling it out to a sample of classrooms who receive ST Math. This rollout will be randomly assigned, allowing me to compare versions of the software and their effects on student calibration and performance outcomes.

Ability to exercise control. The importance of accurate metacognitive judgments depends on the student's ability to exercise control to improve his/her learning and performance. Within the current version of ST Math, students can direct their attention and can replay levels that they have already passed. Additional control may be necessary for the students to take the kind of ownership of their learning assumed within systems of SRL. One potential strength of ST Math is the visual feedback offered—within the games, students can see the consequences of their selections represented as visual models of mathematics. It is not known how much attention students pay to these representations, nor do students have the ability to explore different selection consequences without risking their score and their progress within the levels of the games. I have been collaborating with MIND to include a “sandbox mode” in certain games. This mode will allow students to exit the typical game-play structure to explore the ramifications

of answer choices within a low-stakes environment. By allowing this additional control over their own learning, I can explore how accuracy in metacognitive judgments relates to behavioral indications of control. By varying the objectives that include this sandbox mode, I can test how these additional control allowances influence the relation between calibration and performance.

I also plan to explore the relation between calibration and control within the current version of ST Math. To examine how students adjust attention during game-play, I will use eye-tracking to measure gaze patterns and analyze their association with levels of certainty across objectives and across different types of material within objectives. Student ability to both monitor and use monitoring to adjust attention during control may vary depending on attentional resources. I have measures of these attentional resources (as working memory and executive functions) for a sample of students within this dissertation. Examining how these resources moderate both the association between calibration and performance and the effect of ST Math on calibration will provide information on whether and how certain cognitive resources are used in metacognition and SRL. Although options for control are limited within the current version of ST Math, students are able to exercise control through replaying levels. Analyzing the association between student level replays and measures of calibration can provide information on whether replays are being made when students note uncertainty and if these replays have positive associations with performance.

Outside of ST Math. Although ST Math provides a rich context for research on calibration and SRL, it is also limiting: students do not have many options for social contact within ST Math, teachers play a limited role, and the subject-matter is constrained to the specific objectives and problem formats of the game. Examining contexts outside of ST Math to build upon the research in this dissertation will provide a more generalizable picture of calibration and

SRL and will allow for finer manipulations to answer some of the questions raised by the results. In particular, I plan to examine how accuracy of confidence judgments across grain sizes is related to performance, how students characterize tasks and cues, and how metacognitive judgments are related to control processes in the classroom.

Calibration across grain sizes. As was noted above, one theme to emerge from the three studies within this dissertation was the complicated divisions between the Task x Person and Person levels within metacognition. Although it is seen as desirable for students to have a slightly overconfident view of their ability to complete a task (see Bandura, 1986), it may be that the desired level of confidence varies depending on the grain size examined. For example, the most positive self-views may be adaptive at the broadest level, such as general self-esteem, whereas accurate calibration or even underconfidence may be adaptive at the item level. I will begin to examine these differing relations by measuring calibration along a continuum of grain sizes and relating these measures to performance. As a first step, I have data on broader mathematics self-efficacy for the sample of students included in this dissertation and can answer questions about the stability of their calibration across grain sizes as well as which level of measurement provides the most predictive power towards achievement.

Student characterization of tasks and cues. Grain size can be defined by researchers (e.g., item, task, subject), but it can also be defined by the students. As students categorize types of tasks, they will draw on different prior experiences depending on the overlap between these experiences and their defined categories. A student who incorrectly identifies the type of problem will draw upon prior knowledge and experiences that may not be useful when forming their metacognitive judgments. Qualitative research is needed to understand how students categorize problem types and how they make determinations of confidence and uncertainty.

Some of this work has been conducted using contexts of knowledge recall (e.g., Dinsmore & Parkinson, 2013), but processes and student interpretation of processes in elementary mathematics may be substantially different than those in recall tasks.

Calibration and control in the classroom. Although MIND is implementing changes to ST Math to allow for greater control within the software, much of student learning takes place in the classroom where the kinds of control activities in which students can engage are less structured. One option available in most classes is seeking help—help seeking has long been considered an important option for control within systems of SRL (Nelson-LeGall, 1981). As part of the data collection for the larger ST Math study, I collected data on student classroom help-seeking behaviors in both mathematics and English/Language Arts along with classroom quiz answers, including calibration ratings. Exploring the interplay between monitoring and this specific aspect of control can help us understand whether students who recognize a need for help actually seek it and whether those who seek help do so as part of conscious regulation or unconscious regulation (see Efklides, 2008).

References

- Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3.
doi:10.1016/j.learninstruc.2012.10.003
- Bandura, Albert. (1986). *Social foundations of thought and action : a social cognitive theory*. Englewood Cliffs N.J.: Prentice-Hall.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology, 64*(1), 417–444. doi:10.1146/annurev-psych-113011-143823
- Chang, A., Rutherford, T., & Farkas, G. (2014, April). *I can do it!: Expectancy as a mediator of the ST Math effect on math achievement*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252.
doi:10.1016/j.learninstruc.2012.01.003
- De Corte, E., Verschaffel, L., & Op ’T Eynde, P. (2005). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation*. Elsevier.
- Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students’ explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. doi:10.1016/j.learninstruc.2012.06.001

- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287.
doi:10.1027/1016-9040.13.4.277
- Efklides, A. (2011). Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist, 46*(1), 6–25.
doi:10.1080/00461520.2011.538645
- Feurman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice, 14*(5), 930–933. doi:10.1111/j.1365-2753.2008.00984.x
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. doi:10.1037/0003-066X.34.10.906
- Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling Uncertain and Acting on It During the Preschool Years: A Metacognitive Approach. *Child Development Perspectives, 7*(3), 160–165. doi:10.1111/cdep.12035
- Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*(4), 559–574. doi:10.1016/S0191-8869(02)00028-4
- Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*(3), 321–341.
doi:10.1002/acp.705
- Nelson-LeGall, S. (1981). Help-seeking: An understudied problem-solving skill in children. *Developmental Review, 1*(3), 224–246. doi:10.1016/0273-2297(81)90019-8

- Rutherford, T., Hinga, B., Chang, A., Conley, A. M., & Martinez, M. E. (2011, August). *The effect of ST Math software on standardized test scores via improvement in mathematics expectancy*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.
- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57.
doi:10.1016/j.learninstruc.2012.08.007
- Winne, P. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466–488.
- Zhao, Q., & Linderholm, T. (2008). Adult Metacomprehension: Judgment Processes and Accuracy Constraints. *Educational Psychology Review, 20*(2), 191–206.
doi:10.1007/s10648-008-9073-8
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. doi:10.3102/0002831207312909

Appendix A
Supplementary Tables for Study 1

Table 1
Associations between Number of Questions Completed and Student Variables

	Grade 2 N=878	Grade 3 N=789	Grade 4 N=1,475	Grade 5 N=1,004
Male	11.785* (4.750)	12.214* (5.209)	9.199** (3.562)	7.545 (4.077)
White	-2.396 (9.604)	14.087 (10.407)	4.743 (8.037)	14.781 (8.681)
Asian	23.471* (11.822)	45.653** (14.773)	38.253*** (11.039)	44.928** (15.122)
Other Ethnicity	10.958 (12.338)	9.758 (15.499)	-3.827 (9.926)	3.394 (13.085)
Eng. Lang Learner	-14.539* (5.814)	-7.278 (6.791)	-11.981** (4.567)	-13.437** (4.463)
Free/Reduced Lunch	-41.625*** (6.980)	-30.850*** (7.209)	-19.065*** (5.261)	-5.445 (6.149)
Constant	200.775*** (7.672)	200.129*** (8.300)	176.249*** (5.689)	145.557*** (6.241)
R2	0.100	0.076	0.043	0.036

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

Standard errors in parentheses. Reference groups are Hispanic students, non-ELL, female, and not eligible for free/reduced lunch.

Table 2

Percent of Students with Zeroes in Quadrant A by Amount of Quiz Questions Sampled

	2nd	3rd	4th	5th
25 Qs	0.01% (.003-.02%)	0.29% (.26-.32)	0.03% (.02-.04)	0.02% (.01-.03)
50 Qs	0.00% N/A	0.06% (.04-.08)	0.00% N/A	0.00% N/A
75 Qs	0.00% N/A	0.03% (.02-.04)	0.00% N/A	0.00% N/A
100 Qs	0.00% N/A	0.01% (.01-.02)	0.00% N/A	0.00% N/A
150 Qs	0.00% N/A	0.00% N/A	0.00% N/A	0.00% N/A
All Qs	0.00%	0.00%	0.00%	0.00%
Ave No. Qs	244.72	258.36	244.05	220.24

Table 3

Percent of Students with Zeroes in Quadrant B by Amount of Quiz Questions Sampled

	2nd	3rd	4th	5th
25 Qs	3.94%	4.37%	3.69%	8.17%
	(3.81-4.08%)	(4.22-4.51)	(3.58-3.80)	(8.04-8.30)
50 Qs	0.99%	1.22%	0.68%	5.90%
	(.93-1.06%)	(1.16-1.29)	(.63-.73)	(5.82-5.97)
75 Qs	0.39%	0.67%	0.19%	5.40%
	(.34-.43)	(.62-.72)	(.17-.22)	(5.34-5.46)
100 Qs	0.15%	0.41%	0.06%	5.16%
	(.13-.18)	(.38-.45)	(.05-.08)	(5.10-5.21)
150 Qs	0.03%	0.23%	0.01%	4.87%
	(.02-.04)	(.19-.26)	(.001-.01)	(4.82-4.92)
All Qs	0.00%	0.00%	0.00%	4.37%
Ave No. Qs	244.72	258.36	244.05	220.24

Table 4

Percent of Students with Zeroes in Quadrant C by Amount of Quiz Questions Sampled

	2nd	3rd	4th	5th
25 Qs	40.20%	42.68%	34.32%	44.24%
	(39.96-40.44%)	(42.42-42.94)	(34.12-34.52)	(43.93-44.55)
50 Qs	29.38%	29.81%	23.80%	32.34%
	(29.19-29.58%)	(29.57-30.04)	(23.63-23.98)	(32.08-32.60)
75 Qs	24.21%	23.95%	19.20%	27.05%
	(24.03-24.39)	(23.77-24.12)	(19.10-19.34)	(26.84-27.26)
100 Qs	21.15%	20.45%	16.30%	23.98%
	(20.99-21.32)	(20.29-20.62)	(16.17-16.43)	(23.80-24.16)
150 Qs	16.75%	16.01%	12.78%	19.79%
	(16.61-16.89)	(15.88-16.15)	(12.67-12.89)	(19.64-19.95)
All Qs	12.28%	11.39%	9.03%	16.59%
Ave No. Qs	244.72	258.36	244.05	220.24

Table 5

Percent of Students with Zeroes in Quadrant D by Amount of Quiz Questions Sampled

	2nd	3rd	4th	5th
25 Qs	31.38%	33.16%	25.01%	39.29%
	(31.15-31.62%)	(32.93-33.39)	(24.85-25.28)	(39.04-39.55)
50 Qs	21.09%	21.73%	16.48%	29.97%
	(20.89-21.29%)	(21.52-21.95)	(16.33-16.62)	(29.76-30.18)
75 Qs	16.29%	16.65%	12.66%	25.22%
	(16.11-16.46)	(16.47-16.84)	(12.54-12.79)	(25.02-25.42)
100 Qs	13.24%	13.65%	10.43%	22.40%
	(13.09-13.40)	(13.47-13.83)	(10.33-10.52)	(22.23-22.58)
150 Qs	9.22%	9.93%	7.67%	18.72%
	(9.09-9.35)	(9.8-10.06)	(7.58-7.76)	(18.59-18.86)
All Qs	5.26%	5.70%	5.07%	15.72%
Ave No. Qs	244.72	258.36	244.05	220.24

Table 6a

Ten Measures of Calibration Calculated Based on Student Pretest Measures Only, Second and Third Grade

	2nd Grade (N=915)						3rd Grade (N=812)						
	N=4,278	Mean	SD	Min	Max	Count	% Valid	Mean	SD	Min	Max	Count	% Valid
Sensitivity		0.86	0.18	0.02	1.00	914	99.89%	0.86	0.18	0.00	1.00	812	100.00%
Specificity		0.31	0.28	0.00	1.00	906	99.02%	0.29	0.25	0.00	0.99	811	99.88%
Simple Match		0.63	0.13	0.00	1.00	915	100.00%	0.63	0.14	0.17	1.00	812	100.00%
Gamma		0.50	0.48	-1.00	1.00	791	86.45%	0.49	0.48	-1.00	1.00	738	90.89%
G Index		0.26	0.27	-1.00	1.00	915	100.00%	0.27	0.28	-0.67	1.00	812	100.00%
Odds Ratio		5.17	6.68	0.00	72.00	666	72.79%	5.35	7.45	0.00	96.00	634	78.08%
Kappa		0.16	0.18	-0.50	1.00	906	99.02%	0.16	0.17	-0.36	0.72	811	99.88%
Phi		0.22	0.19	-0.50	1.00	791	86.45%	0.21	0.18	-0.41	0.73	738	90.89%
Sokal Reverse		0.59	0.12	0.00	1.00	915	100.00%	0.59	0.12	0.00	0.91	812	100.00%
Discrimination		0.71	0.54	-0.79	2.47	629	68.74%	0.71	0.53	-0.95	2.39	598	73.65%

Table 6b

Ten Measures of Calibration Calculated Based on Student Pretest Measures Only, Fourth and Fifth Grade

	4th Grade (N=1,521)						5th Grade (N=1,030)					
	Mean	SD	Min	Max	Count	% Valid	Mean	SD	Min	Max	Count	% Valid
Sensitivity	0.83	0.18	0.00	1.00	1521	100.00%	0.85	0.17	0.00	1.00	1030	100.00%
Specificity	0.35	0.28	0.00	1.00	1517	99.67%	0.31	0.26	0.00	1.00	1012	98.64%
Simple Match	0.62	0.13	0.09	1.00	1521	100.00%	0.64	0.13	0.20	1.00	1030	100.00%
Gamma	0.52	0.43	-1.00	1.00	1352	94.22%	0.50	0.46	-1.00	1.00	891	91.38%
G Index	0.23	0.26	-0.83	1.00	1521	100.00%	0.28	0.27	-0.60	1.00	1030	100.00%
Odds Ratio	5.11	5.97	0.00	87.43	1185	85.48%	5.17	11.22	0.00	227.50	765	82.46%
Kappa	0.18	0.17	-0.67	0.67	1517	99.67%	0.17	0.18	-0.67	1.00	1022	99.42%
Phi	0.23	0.18	-0.80	0.71	1352	94.22%	0.22	0.18	-0.67	1.00	891	91.38%
Sokal Reverse	0.61	0.11	0.00	0.96	1521	100.00%	0.59	0.13	0.00	0.89	1030	100.00%
Discrimination	0.74	0.50	-0.93	2.34	1140	75.00%	0.72	0.48	-0.68	3.03	722	70.10%

Note. Includes all students who have valid pretest data. Ten calibration measures are calculated as in Schraw et al. (2013) based on four quadrants of agreement between accuracy and confidence. % Valid represents the percent of students for whom the given measures is calculable.

Table 7a

Regression of Posttest Accuracy on Pretest Calibration and Accuracy for Ten Measures of Calibration: Diagnostic Efficiency & Agreement Measures

N=3,033	(1)	(2)	(3)	(4)	(5)	(6)	
	Acc. Only	Sensitivity	Specificity	Sensitivity	Specificity	Simple Match	G Index
Measure(s)		0.046*** (0.009)	-0.003 (0.006)	0.098*** (0.014)	0.047*** (0.010)	0.074*** (0.018)	0.037*** (0.009)
Pretest Accuracy	0.818*** (0.012)	0.803*** (0.012)	0.818*** (0.012)	0.789*** (0.012)		0.779*** (0.015)	0.779*** (0.015)
No. Pretest Qs.	0.0005 (0.0004)	0.0004 (0.0004)	0.0005 (0.0004)	0.00050 (0.0004)		0.00050 (0.0004)	0.00050 (0.0004)
No. Posttest Qs	0.0004 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)	0.0003 (0.0004)		0.0004 (0.0004)	0.0004 (0.0004)
Male	-0.014*** (0.003)	-0.016*** (0.003)	-0.014*** (0.003)	-0.015*** (0.003)		-0.014*** (0.003)	-0.014*** (0.003)
Asian	-0.013 (0.009)	-0.013 (0.009)	-0.013 (0.009)	-0.014 (0.009)		-0.014 (0.009)	-0.014 (0.009)
White	-0.004 (0.006)	-0.003 (0.006)	-0.004 (0.006)	-0.004 (0.006)		-0.005 (0.006)	-0.005 (0.006)
Other Race	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)		-0.001 (0.009)	-0.001 (0.009)
ELL	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)		-0.005 (0.004)	-0.005 (0.004)
Free/Reduc Lunch	-0.008 (0.005)	-0.008 (0.005)	-0.008 (0.005)	-0.007 (0.005)		-0.007 (0.005)	-0.007 (0.005)
Grade 2	-0.001 (0.004)	-0.001 (0.004)	-0.001 (0.004)	-0.0001 (0.004)		0.0005 (0.004)	0.0005 (0.004)
Grade 3	-0.010* (0.004)	-0.010* (0.004)	-0.010* (0.004)	-0.008 (0.004)		-0.008 (0.004)	-0.008 (0.004)
Grade 5	0.027*** (0.004)	0.026*** (0.004)	0.027*** (0.004)	0.027*** (0.004)		0.028*** (0.004)	0.028*** (0.004)
Constant	0.147*** (0.009)	0.121*** (0.011)	0.148*** (0.010)	0.066*** (0.015)		0.121*** (0.011)	0.158*** (0.009)
R2	0.697	0.699	0.697	0.702		0.698	0.698

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients. Standard errors in parentheses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch. Sample limited to those students who have non-missing values for each of the ten measures of calibration as described in Schraw et al. (2013).

Table 7b

Regression of Posttest Accuracy on Pretest Calibration and Accuracy for Ten Measures of Calibration: Association, Binary Distance, and Discrimination

N=3,033	(7)	(8)	(9)	(10)	(11)	(12)
	Gamma	Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
Measure(s)	0.028*** (0.005)	0.0004* (0.0002)	0.046*** (0.010)	0.049*** (0.010)	-0.081*** (0.022)	0.017*** (0.003)
Pretest Accuracy	0.798*** (0.012)	0.812*** (0.012)	0.804*** (0.012)	0.803*** (0.012)	0.781*** (0.015)	0.799*** (0.012)
No. Pretest Qs	0.00050 (0.0004)	0.00050 (0.0004)	0.00050 (0.0004)	0.00050 (0.0004)	0.00050 (0.0004)	0.00050 (0.0004)
No. Posttest Qs	0.0004 (0.0004)	0.0004 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)
Male	-0.014*** (0.003)	-0.014*** (0.003)	-0.013*** (0.003)	-0.013*** (0.003)	-0.014*** (0.003)	-0.014*** (0.003)
Asian	-0.014 (0.009)	-0.014 (0.009)	-0.014 (0.009)	-0.014 (0.009)	-0.014 (0.009)	-0.015 (0.009)
White	-0.005 (0.006)	-0.004 (0.006)	-0.005 (0.006)	-0.005 (0.006)	-0.005 (0.006)	-0.005 (0.006)
Other Race	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)	-0.001 (0.009)
ELL	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.005 (0.004)
Free/Reduced Lunch	-0.008 (0.005)	-0.008 (0.005)	-0.007 (0.005)	-0.007 (0.005)	-0.007 (0.005)	-0.007 (0.005)
Grade 2	0.0003 (0.004)	-0.001 (0.004)	0.0001 (0.004)	0.0002 (0.004)	0.0003 (0.004)	0.0002 (0.004)
Grade 3	-0.008 (0.004)	-0.009* (0.004)	-0.008 (0.004)	-0.008 (0.004)	-0.009* (0.004)	-0.008 (0.004)
Grade 5	0.028*** (0.004)	0.027*** (0.004)	0.028*** (0.004)	0.028*** (0.004)	0.028*** (0.004)	0.028*** (0.004)
Constant	0.144*** (0.009)	0.147*** (0.009)	0.143*** (0.009)	0.142*** (0.009)	0.215*** (0.020)	0.145*** (0.009)
R2	0.700	0.697	0.699	0.699	0.698	0.699

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients. Standard errors in parentheses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch. Sample limited to those students who have non-missing values for each of the ten measures of calibration as described in Schraw et al. (2013).

Table 8
Demographic Information and Descriptive Statistics for All Students with Quiz Data

Sample of Students Who Answered > 200 Questions					
	Mean/Percent	SD	Min	Max	Count
Grade 2	21%				4,278
Grade 3	19%				4,278
Grade 4	36%				4,278
Grade 5	24%				4,278
Male	52%				4,145
Asian	3%				4,145
Hispanic	85%				4,145
White	8%				4,145
Other Race	3%				4,145
English Lang Learner	65%				4,144
Free/Reduced Lunch	81%				4,145
ELA CST	336.75	60.19	171	600	3,075
Math CST	373.35	80.29	150	600	3,070
Pretest Quiz Accuracy	0.56	0.14	0.15	0.98	4,278
Posttest Quiz Accuracy	0.67	0.16	0	1	4,278
Total Pretest Questions	77.85	36.68	3	140	4,278
Total Posttest Questions	78.82	36.76	2	140	4,278
N	4,278				

Note. This table presents information for those students who had at least one valid pre and posttest and who were included in the replication of the analysis of calibration measures.

Table 9a

Regression of Posttest Accuracy on Pretest Calibration and Accuracy for Ten Measures of Calibration: Diagnostic Efficiency & Agreement Measures, Adjusted for Missingness

N=4,144	(1)	(2)	(3)	(4)	(5)	(6)	
	Acc. Only	Sensitivity	Specificity	Sensitivity	Specificity	Simple Match	G Index
Measure(s)		0.035*** (0.009)	0.003 (0.006)	0.111*** (0.015)	0.061*** (0.010)	0.080*** (0.018)	0.040*** (0.009)
Pretest Accuracy	0.871*** (0.011)	0.860*** (0.011)	0.871*** (0.011)	0.838*** (0.012)		0.819*** (0.016)	0.819*** (0.016)
No. Pretest Qs	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)		0.0003 (0.0004)	0.0003 (0.0004)
No. Posttest Qs	0.0004 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)		0.0004 (0.0004)	0.0004 (0.0004)
Male	-0.016*** (0.003)	-0.017*** (0.003)	-0.015*** (0.003)	-0.016*** (0.003)		-0.015*** (0.003)	-0.015*** (0.003)
Asian	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)		-0.002 (0.009)	-0.002 (0.009)
White	-0.004 (0.006)	-0.003 (0.006)	-0.004 (0.006)	-0.005 (0.006)		-0.005 (0.006)	-0.005 (0.006)
Other Race	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)		-0.006 (0.008)	-0.006 (0.008)
ELL	-0.008* (0.004)	-0.007* (0.004)	-0.008* (0.004)	-0.007* (0.004)		-0.007* (0.004)	-0.007* (0.004)
Free/Reduced Lunch	-0.006 (0.004)	-0.006 (0.004)	-0.006 (0.004)	-0.005 (0.004)		-0.005 (0.004)	-0.005 (0.004)
Grade 2	0.004 (0.004)	0.004 (0.004)	0.004 (0.004)	0.006 (0.004)		0.006 (0.004)	0.006 (0.004)
Grade 3	-0.007 (0.004)	-0.007 (0.004)	-0.007 (0.004)	-0.005 (0.004)		-0.006 (0.004)	-0.006 (0.004)
Grade 5	0.029*** (0.004)	0.029*** (0.004)	0.029*** (0.004)	0.030*** (0.004)		0.030*** (0.004)	0.030*** (0.004)
Constant	0.130*** (0.008)	0.108*** (0.010)	0.129*** (0.009)	0.037* (0.015)		0.109*** (0.010)	0.149*** (0.009)
R2	0.670	0.672	0.670	0.675		0.672	0.672

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients. Standard errors in parentheses. The reference group comprises students who were in fourth grade, Hispanic, Non-ELL, and not on free lunch.

Regressions on full sample of students who have at least one valid pre and posttest. To eliminate missing values of calibration due to zero quadrants, a "1" was added to each quadrant before calculating the calibration measures.

Table 9b

Regression of Posttest Accuracy on Pretest Calibration and Accuracy for Ten Measures of Calibration: Association, Binary Distance, and Discrimination, Adjusted for Missingness

N=4,144	(7)	(8)	(9)	(10)	(11)	(12)
	Gamma	Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
Measure(s)	0.029*** (0.005)	0.000 (0.000)	0.052*** (0.010)	0.057*** (0.009)	-0.075*** (0.022)	0.018*** (0.003)
Pretest Accuracy	0.841*** (0.012)	0.866*** (0.012)	0.853*** (0.012)	0.849*** (0.012)	0.830*** (0.016)	0.841*** (0.012)
No. Pretest Qs	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)
No. Posttest Qs	0.0004 (0.0004)	0.0004 (0.0004)	0.0003 (0.0004)	0.0003 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)
Male	-0.014*** (0.003)	-0.016*** (0.003)	-0.014*** (0.003)	-0.014*** (0.003)	-0.015*** (0.003)	-0.014*** (0.003)
Asian	-0.002 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)	-0.003 (0.009)
White	-0.005 (0.006)	-0.004 (0.006)	-0.005 (0.006)	-0.005 (0.006)	-0.005 (0.006)	-0.005 (0.006)
Other Race	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)	-0.006 (0.008)
ELL	-0.008* (0.004)	-0.008* (0.004)	-0.008* (0.004)	-0.008* (0.004)	-0.007* (0.004)	-0.008* (0.004)
Free/Reduced Lunch	-0.005 (0.004)	-0.006 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.006 (0.004)	-0.005 (0.004)
Grade 2	0.005 (0.004)	0.004 (0.004)	0.006 (0.004)	0.006 (0.004)	0.005 (0.004)	0.006 (0.004)
Grade 3	-0.005 (0.004)	-0.007 (0.004)	-0.005 (0.004)	-0.005 (0.004)	-0.006 (0.004)	-0.005 (0.004)
Grade 5	0.030*** (0.004)	0.029*** (0.004)	0.030*** (0.004)	0.030*** (0.004)	0.030*** (0.004)	0.030*** (0.004)
Constant	0.135*** (0.008)	0.131*** (0.009)	0.130*** (0.008)	0.131*** (0.008)	0.198*** (0.022)	0.135*** (0.008)
R2	0.673	0.670	0.673	0.673	0.671	0.672

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients. Standard errors in parentheses. The reference group comprises students who were in fourth grade, Hispanic, Non-ELL, and not on free lunch. Regressions on full sample of students who have at least one valid pre and posttest. To eliminate missing values of calibration due to zero quadrants, a “1” was added to each quadrant before calculating the calibration measures.

Appendix B
Supplementary Figures and Tables for Study 2

Table 1
List of ST Math Objectives for 2010 Curriculum, Divided by Grade

Grade 2	Grade 3	Grade 4	Grade 5
Addition and Subtraction Facts	Place Value	Place Value	Place Value
Addition and Subtraction Relationships	Ordering and Comparing Whole Numbers	Expanded Notation	Decimal Addition and Subtraction
Place Value to 200	Expanded Notation	Ordering and Comparing Whole Numbers	Exponents
Comparing Numbers to 200	Addition to 10,000	Addition and Subtraction	Variables
Classifying Shapes	Subtraction to 10,000	Factorization and Prime Numbers	Lines and Angles
Addition and Subtraction, Sums to 200	2D Shapes	Variables and Unknowns	Factorization
Measurement	3D Shapes	Lines and Angles	Fraction Concepts
Fractions to a Whole	Lines and Angles	Shapes and Attributes	Fraction Addition and Subtraction
Telling Time	Multiplication Concepts	Fraction Concepts	Fraction Addition and Subtraction LI
Elapsed Time	Multiplication Facts	Fraction Addition and Subtraction	Relationships of Decimals, Fractions, Percents
Time Relationships	Division	Fraction Addition and Subtraction LI	Ordered Pairs and Graphing
Place Value to 1000	Algebraic Expressions and Equations	Decimal Fraction Relationships and Equivalence	Mean Median Mode
Expanded Forms	Functional Relationships	Decimals and Fractions PV	Whole Number Operations
Range and Mode	Fraction Concepts	Decimal Operations and Money	Fraction Multiplication and Division
Patterns and Functions	Fraction and Decimal Equivalence	Whole Number Multiplication and Division	Area
More Fraction Concepts	Fraction Addition and Subtraction	Using Parentheses	Volume
Using Data, Charts and Graphs	Fraction Addition and Subtraction LI	Equations	Negative Number Operations
Comparing Numbers to 1000	Multiplication of Multi Digits	Integers	Parentheses
Comparing Fractions	Measurement	Area and Perimeter	Linear Functions and Equations
Money and Decimals	Area, Perimeter and Volume	Using Data and Graphs	Using Data and Graphs
Addition and Subtraction to 1000	Money	Coordinate Grids and Ordered Pairs	Temperature and Capacity
Multiplication Concepts	Outcomes	Symmetry	
Multiplication Facts	Using Data and Graphs	Median Mode	
Division Concepts	Temperature and Capacity	Outcomes	
Temperature and Capacity		Temperature and Capacity	

4:23 **3**

What is 2,853 rounded to the nearest hundred?

(A) 2,000

(B) 3,000


(C) 2,900


(D) 2,950


Confidence Level:


4:19 **4**

Which angle is a right angle?

(A) 


(B) 

(C) 


(D) 


Confidence Level:


4:22 **3**




What fractional part of the rectangle is equal to $\frac{4}{12}$?

(A) 

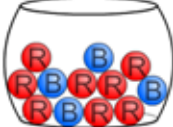
(B) 

(C) 

(D) 

Confidence Level:

4:15 **3**



If you pick one marble from the jar, is it certain, likely, unlikely, or impossible that the marble you pick is red?

(A) Certain

(B) Likely, but not certain

(C) Unlikely, but not impossible

(D) Impossible

Confidence Level:

(a)
(b)

(c)
(d)

Figure 2. Quiz examples from third grade curriculum. (a) Objective 1, Place Value. (b) Objective 8, Lines and Angles. (c) Objective 14, Fraction Concepts. (d) Objective 22, Outcomes.

Table 2

Correlations between Quiz Measures and Student Demographic Characteristics

	Sensitivity	Specificity	Pretest Acc.	Posttest Acc.
Grade 2	0.034*	-0.033*	0.107***	0.068***
Grade 3	0.062***	-0.054***	0.057***	0.007
Grade 4	-0.103***	0.068***	-0.135***	-0.109***
Grade 5	0.026	0.004	-0.003	0.053***
Male	0.090***	-0.169***	-0.116***	-0.117***
Asian	0.046**	-0.012	0.096***	0.090***
Hispanic	-0.051**	-0.035*	-0.194***	-0.186***
White	0.016	0.054***	0.143***	0.139***
Other Ethnicity	0.032*	-0.003	0.069***	0.067***
Eng. Lang. Learner	-0.056***	-0.019	-0.188***	-0.199***
Free/Reduced Lunch	-0.050**	-0.026	-0.168***	-0.183***

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

Table 3

Model Fit Statistics for Hierarchical Linear Models

	(1)	(2)	(3)	(4)	(5)	(6)
	Non-HLM	Unconditional	Covariates Only	Add Pretest Acc.	Full conditional	With Interactions
Total Variance	0.0806824	0.0812649	0.0791443	0.0572135	0.0568569	0.0567945
L2 (Student)		0.0179063	0.0157771	0.0025576	0.0025381	0.0024973
Residual (Quiz)		0.0633586	0.0633672	0.0546559	0.0543188	0.0542972
<i>Percent of residual variance by level</i>						
L2 (Student)		22.03%	19.93%	4.47%	4.46%	4.40%
Residual (Quiz)		77.97%	80.07%	95.53%	95.54%	95.60%
<i>Percent Reduction of variance by level from prior model</i>						
L2 (Student)			11.89%	83.79%	0.76%	1.61%
Residual (Quiz)			-0.01%	13.75%	0.62%	0.04%
Overall			2.61%	27.71%	0.62%	0.11%
<i>Percent Reduction of variance from Unconditional model</i>						
Overall			2.61%	29.60%	30.04%	30.11%
<i>Test of Statistical Significance for Model Changes</i>						
Deviance	18264.413	10602.33	10232.773	-1950.7263	-2305.3158	-2351.6951
Parameter change		1	8	2	4	12
Deviance Change		7662.083	369.557	12183.4993	354.5895	46.3793
Needed change		3.841	15.507	5.991	9.488	21.026
Improvement		Yes	Yes	Yes	Yes	Yes

Note. Model fit statistics for HLM analysis of within and between student associations between calibration and improvements in quiz accuracy, including a final model with grade-level interactions (Model 6).

Table 4

Grade & Demographic Information for 2011 Replication Sample

	Total Sample		Analysis Sample	
	Percent	Count	Percent	Count
Grade 2	26%	6,254	26%	6,091
Grade 3	24%	6,254	24%	6,091
Grade 4	25%	6,254	25%	6,091
Grade 5	25%	6,254	25%	6,091
Male	52%	6,254	52%	6,091
Asian	3%	6,254	3%	6,091
Hispanic	74%	6,254	74%	6,091
White	20%	6,254	20%	6,091
Other Ethnicity	3%	6,254	3%	6,091
English Language Learner	62%	6,252	62%	6,091
Nat'l Free/Reduced Lunch	84%	6,252	84%	6,091
	N	6,254		6,091

Note. Total Sample includes all students in second through fifth grade in the study schools who began at least one objective within ST Math in the 2011-2012 school year. The analysis sample is limited to those students who had complete demographic information and completed at least two complete objectives (pre and posttest).

Table 5

Quiz Accuracy and Calibration Measures, by Grade, 2011 Replication Sample

Observation/Objective-Level Quiz Descriptives (N=84,308)								
	Grade 2		Grade 3		Grade 4		Grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest Accuracy	0.63	0.29	0.67	0.27	0.55	0.29	0.58	0.30
Pretest Sensitivity	0.85	0.28	0.84	0.29	0.76	0.35	0.78	0.34
Pretest Specificity	0.30	0.34	0.36	0.37	0.38	0.39	0.37	0.39
Posttest Accuracy	0.72	0.26	0.78	0.23	0.69	0.28	0.71	0.27
N (Observations)	21,315		22,149		21,514		19,330	

Student-Level Quiz Descriptive Statistics (N=6,091)								
	Grade 2		Grade 3		Grade 4		Grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest Accuracy	0.61 ^a	0.13	0.65 ^a	0.15	0.53 ^b	0.14	0.56	0.15
Pretest Sensitivity	0.85 ^a	0.15	0.83	0.19	0.76 ^b	0.20	0.76 ^b	0.21
Pretest Specificity	0.29 ^b	0.18	0.36 ^a	0.22	0.37	0.24	0.38 ^a	0.24
Posttest Accuracy	0.70 ^a	0.12	0.76 ^a	0.13	0.66 ^a	0.15	0.68	0.16
N (Students)	1,604		1,475		1,515		1,497	

Note. Data from analysis sample presented with objective data nested within students. Curricular and quiz content differs across grades.

^a2011 sample has statistically significantly higher value ($p < .05$)

^b2010 sample has statistically significantly higher value ($p < .05$)

Table 6

Model Fit Statistics for Hierarchical Linear Models, 2011 Replication Sample

	(1) Non-HLM	(2) Unconditional	(3) Covariates Only	(4) Add Pretest Acc.	(5) Full conditional	(6) With Interactions
Total Variance	0.0695204	0.0702481	0.0671849	0.0513422	0.0511245	0.05106
L2 (Student)		0.0135398	0.0104316	0.0024068	0.0024092	0.0023845
Residual (Quiz)		0.0567083	0.0567533	0.0489354	0.0487153	0.0486755
<i>Percent of residual variance by level</i>						
L2 (Student)		19.27%	15.53%	4.69%	4.71%	4.67%
Residual (Quiz)		80.73%	84.47%	95.31%	95.29%	95.33%
<i>Percent Reduction of variance by level from prior model</i>						
L2 (Student)			22.96%	76.93%	-0.10%	1.03%
Residual (Quiz)			-0.08%	13.78%	0.45%	0.08%
Overall			4.36%	23.58%	0.42%	0.13%
<i>Percent Reduction of variance from Unconditional model</i>						
Overall			4.36%	26.91%	27.22%	27.31%
<i>Test of Statistical Significance for Model Changes</i>						
Deviance	14479.252	5729.3975	4679.5332	-12008.768	-12374.087	-12464.53
Parameter change		1	8	2	4	12
Deviance Change		8749.8545	1049.8643	16688.3012	365.319	90.443
Needed change		3.841	15.507	5.991	9.488	21.026
Improvement		Yes	Yes	Yes	Yes	Yes

Table 7
Results from Hierarchical Regressions of Post-test Accuracy on Pre-test Accuracy, Calibration, & Covariates, 2011 Replication Sample

<i>Fixed Parameters</i>	B	SE	<i>p</i>	B	SE	<i>p</i>
Level 1						
Sensitivity	0.061	0.003	<.0001	0.064	0.006	<.0001
Specificity	0.010	0.003	<.0001	-0.004	0.005	0.483
Pretest Accuracy	0.319	0.003	<.0001	0.318	0.003	<.0001
GR2*Sensitivity				0.009	0.009	0.322
GR3*Sensitivity				-0.020	0.009	0.024
GR5*Sensitivity				-0.004	0.008	0.617
GR2*Specificity				0.043	0.008	<.0001
GR3*Specificity				-0.008	0.007	0.296
GR5*Specificity				0.023	0.008	0.003
Level 2						
Sensitivity	0.047	0.011	<.0001	0.058	0.019	0.002
Specificity	0.039	0.009	<.0001	0.055	0.016	<.0001
Pretest Accuracy	0.693	0.009	<.0001	0.693	0.009	<.0001
Grade 2	-0.020	0.003	<.0001	0.007	0.028	0.795
Grade 3	0.009	0.003	0.004	0.085	0.026	0.001
Grade 5	-0.001	0.003	0.671	-0.054	0.026	0.042
ELL	-0.016	0.002	<.0001	-0.016	0.002	<.0001
Male	0.009	0.002	<.0001	0.009	0.002	<.0001
Asian	0.022	0.006	<.0001	0.021	0.006	<.0001
White	0.010	0.003	<.0001	0.010	0.003	<.0001
Other Ethnic	0.013	0.006	0.027	0.012	0.006	0.033
Free/Reduced Lunch	-0.006	0.003	0.059	-0.006	0.003	0.045
GR2*Sensitivity				-0.026	0.027	0.342
GR3*Sensitivity				-0.063	0.025	0.01
GR5*Sensitivity				0.053	0.025	0.034
GR2*Specificity				-0.016	0.022	0.463
GR3*Specificity				-0.067	0.021	0.001
GR5*Specificity				0.030	0.021	0.163
Intercept	0.259	0.011	<.0001	0.245	0.019	<.0001
<i>Random Parameters</i>						
Between	0.002	0.0001		0.002	0.0001	
Residual	0.049	0.0002		0.049	0.0002	
<i>% Variance Explained</i>						
L2	0.822			0.824		
L1	0.141			0.142		

Note. Unstandardized regression coefficients. Level 1 variables are group-mean centered around student means. Level 2 quiz variables represent student means. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch.

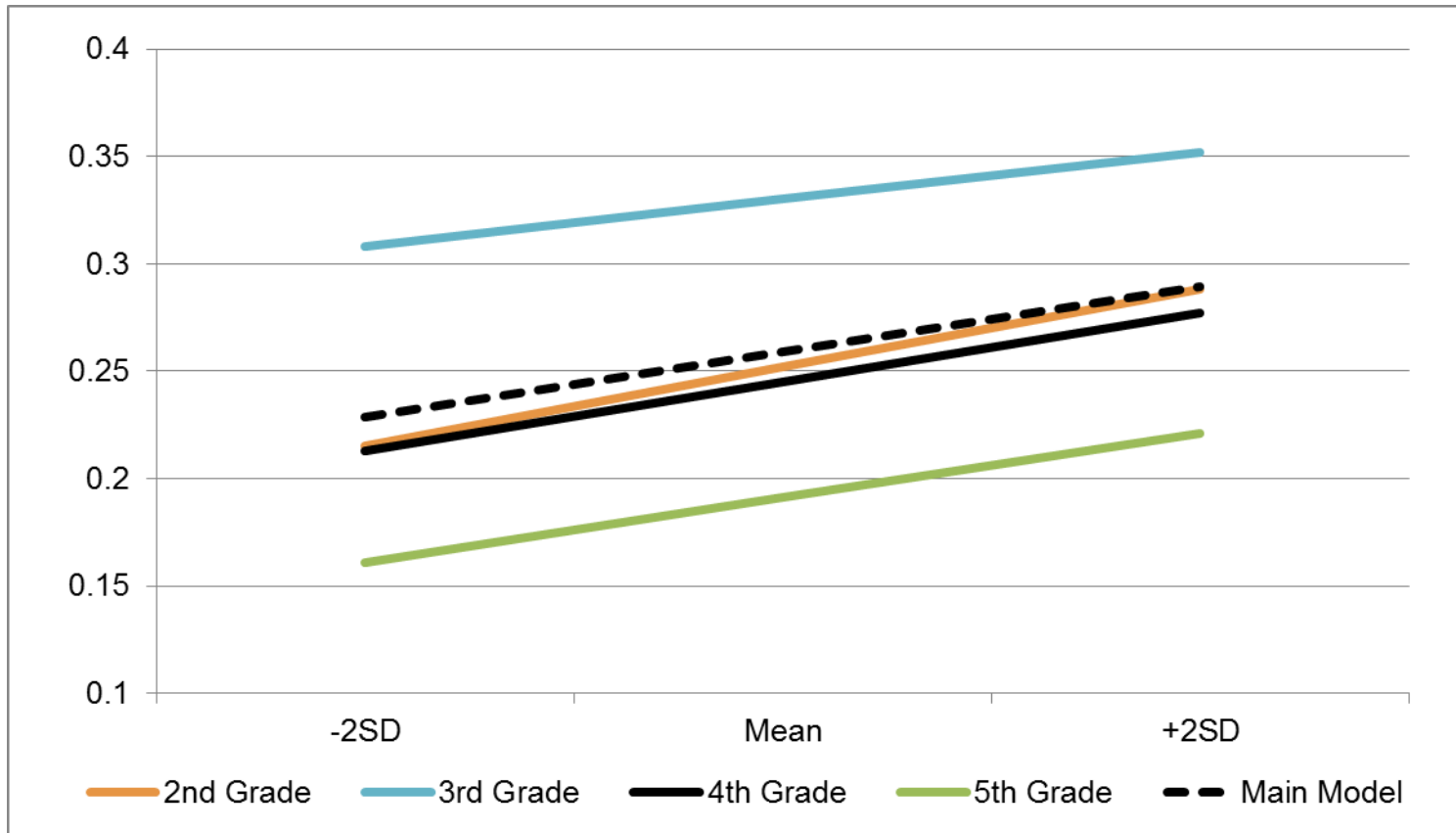


Figure 2. Slopes for Sensitivity by grade-level from 2011 replication sample interaction model on right-hand side of Appendix B, Table 7 compared with slope from non-interaction model (Main Model, dashed line) on left-hand side of Appendix B, Table 7.

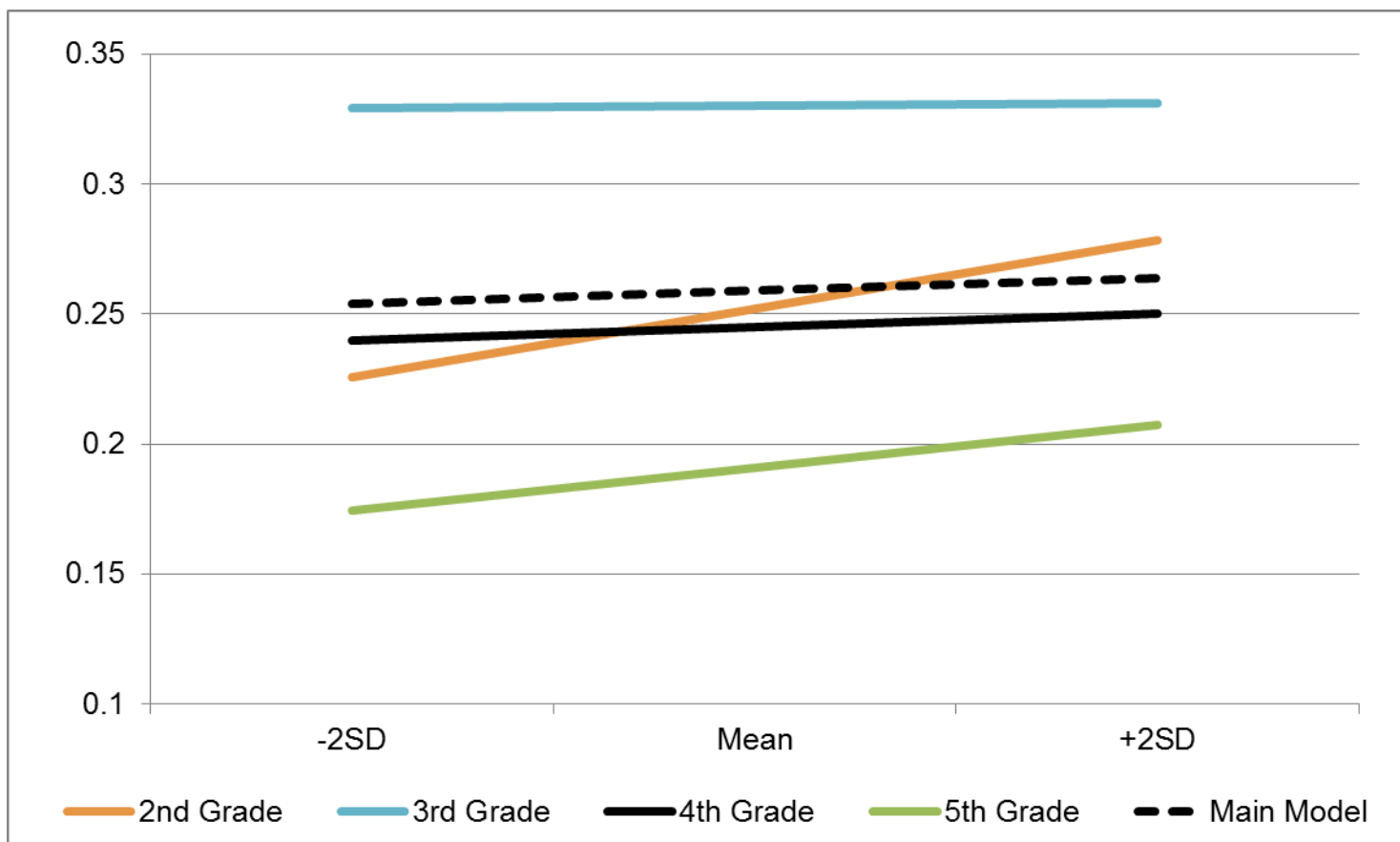


Figure 3. Slopes for Specificity by grade-level from 2011 replication sample interaction model on right-hand side of Appendix B, Table 7 compared with slope from non-interaction model (Main Model, dashed line) on left-hand side of Appendix B, Table 7.

Appendix C
Supplementary Tables for Study 3

Table 1a
Percent of Students Completing Each ST Math Objective, Third Grade

	All Objectives		First Objective		First 3 Objectives	
	ETG	LTG	ETG	LTG	ETG	LTG
1 Place Value to 10,000	98%	97%	98%	97%	98%	97%
2 Ordering and Comparing Whole Numbers	98%	94%	<1%	<1%	98%	94%
3 Addition and Subtraction to 1,000	90%	97%	<1%	<1%	90%	97%
4 Lines and Angles	76%	68%	<1%	<1%	4%	3%
5 2D Shapes	78%	66%	<1%	<1%	2%	<1%
6 3D Shapes	78%	66%			3%	<1%
7 Multiplication Concepts	93%	91%	<1%		<1%	2%
8 Division	77%	79%				
9 Algebraic Expressions and Equations	76%	68%			<1%	2%
10 Functional Relationships	77%	77%				<1%
11 Fraction Concepts	71%	70%	<1%		<1%	
12 Fraction Addition and Subtraction	66%	64%				
13 Money and Decimals	64%	60%			<1%	
14 Measurement	61%	48%			<1%	
15 Area, Perimeter and Volume	62%	46%	<1%		<1%	
16 Addition and Subtraction to 10,000	66%	50%			<1%	
17 Multiplication Facts	65%	62%			<1%	<1%
18 Multiplication of Multi Digits	71%	77%				<1%
19 Fraction and Decimal Equivalence	51%	49%				
20 Outcomes	47%	45%				
21 Using Data and Graphs	52%	43%				<1%
22 Temperature and Capacity	53%	41%				
23 Addition and Subtraction Relationships	12%	13%				
N	649	630	649	630	649	630

Table 1b

Percent of Students Completing Each ST Math Objective, Fourth Grade

	All Objectives		First Objective		First 3 Objectives	
	ETG	LTG	ETG	LTG	ETG	LTG
1 Symmetry	64%	53%	64%	53%	64%	53%
2 Place Value to 1 Million	99%	97%	36%	46%	99%	97%
3 Ordering and Comparing Whole Numbers	98%	99%	<1%	<1%	98%	99%
4 Whole Number Addition and Subtraction	97%	94%			33%	44%
5 Whole Number Multiplication and Division	89%	94%		<1%	1%	4%
6 Fraction Concepts	78%	83%	<1%		2%	<1%
7 Factorization and Prime Numbers	81%	89%				<1%
8 Integers	85%	70%			<1%	<1%
9 Variables and Equations	85%	78%				
10 Input Output	76%	67%				
11 Using Parentheses	78%	69%				
12 Decimals and Fractions	63%	76%				
13 Decimal Operations and Money	65%	72%				
14 Shapes and Attributes	57%	57%			<1%	
15 Lines and Angles	58%	57%			<1%	
16 Area and Perimeter	46%	44%				
17 Graphing on Coordinate Grids	55%	54%			<1%	
18 Median Mode	48%	45%				
19 Using Data and Graphs	50%	44%				
20 Outcomes	39%	40%				
21 Temperature and Capacity	42%	48%				
22 Fraction Addition and Subtraction	4%	11%				
23 Addition and Subtraction Relationships	4%	8%				
N	623	723	623	723	623	723

Table 2
Effect of Early Treatment Group on Calibration for Place Value

	(1)	(2)	(3)	(4)	(5)
N=2,560	Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	-0.054*** (0.009)	0.011 (0.013)	-0.047*** (0.008)	-0.089*** (0.019)	-0.240*** (0.057)
Pretest Accuracy	0.119*** (0.017)	-0.008 (0.025)	0.167*** (0.017)	0.279*** (0.041)	0.701*** (0.115)
Grade 4	0.027* (0.009)	0.056*** (0.013)	0.085*** (0.009)	0.152*** (0.020)	0.473*** (0.060)
Male	0.010 (0.008)	-0.027* (0.012)	-0.002 (0.009)	-0.012 (0.017)	-0.069 (0.056)
Asian	0.027 (0.018)	0.033 (0.044)	0.026 (0.016)	0.060 (0.037)	0.304* (0.141)
White	-0.007 (0.012)	0.004 (0.014)	-0.003 (0.006)	-0.003 (0.017)	-0.046 (0.048)
Other Ethnicity	0.009 (0.020)	-0.027 (0.022)	0.015 (0.018)	0.004 (0.047)	-0.108 (0.128)
Eng. Lang. Learner	-0.014 (0.012)	0.005 (0.012)	-0.015 (0.010)	-0.023 (0.027)	-0.043 (0.076)
Free/Reduced Lunch	-0.004 (0.011)	0.007 (0.016)	0.007 (0.009)	0.016 (0.020)	0.026 (0.064)
Math CST 2011	0.0002* (0.0001)	0.0001 (0.0001)	0.0004*** (0.0001)	0.001*** (0.0002)	0.001*** (0.001)
ELA CST 2011	0.0001 (0.0001)	0.0003 (0.0002)	0.001*** (0.0001)	0.001** (0.0002)	0.003*** (0.001)
Constant	0.748*** (0.050)	0.203** (0.053)	0.371*** (0.037)	-0.133 (0.081)	-0.480 (0.230)
R2	0.065	0.015	0.186	0.105	0.079

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) regression coefficients presented. Standard errors are in parentheses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to those who had data on the Place Value objective, the first objective for all third graders and first or second objective for all fourth graders. Standard errors clustered on school (N=18).

Table 3

Effect of Early Treatment Group on Calibration for First Three Objectives Encountered in ST Math

	(1)	(2)	(3)	(4)	(5)
N=2,624	Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	-0.055*** (0.010)	0.019 (0.010)	-0.044*** (0.007)	-0.079*** (0.016)	-0.216*** (0.050)
Pretest Accuracy	0.186*** (0.033)	0.032 (0.041)	0.281*** (0.028)	0.540*** (0.050)	1.337*** (0.168)
Grade 4	0.007 (0.012)	-0.002 (0.013)	0.022** (0.007)	0.045** (0.012)	0.059 (0.041)
Male	0.032*** (0.007)	-0.041*** (0.006)	0.004 (0.004)	0.006 (0.009)	-0.030 (0.027)
Asian	0.016 (0.024)	0.019 (0.031)	0.017 (0.011)	0.050* (0.017)	0.187*** (0.044)
White	-0.010 (0.013)	0.030** (0.008)	0.022** (0.007)	0.038 (0.019)	0.107 (0.065)
Other Ethnicity	-0.008 (0.018)	-0.026 (0.020)	-0.001 (0.015)	-0.014 (0.036)	-0.125 (0.099)
Eng. Lang. Learner	-0.014 (0.009)	0.005 (0.009)	-0.014** (0.005)	-0.023 (0.012)	-0.049 (0.033)
Free/Reduced Lunch	-0.006 (0.008)	0.006 (0.012)	0.005 (0.007)	0.007 (0.015)	0.009 (0.049)
Math CST 2011	0.0001 (0.0001)	0.00001 (0.0001)	0.0002** (0.0001)	0.0004** (0.0001)	0.001 (0.001)
ELA CST 2011	0.0001 (0.0001)	0.0002 (0.0001)	0.0003*** (0.0001)	0.001* (0.0002)	0.002** (0.001)
Constant	0.582*** (0.122)	0.186 (0.102)	0.104 (0.095)	-0.580** (0.193)	-1.404** (0.440)
R2	0.130	0.046	0.315	0.202	0.156

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) regression coefficients presented. Standard errors are in parentheses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to those who had data on at least three objectives, the specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 4

Effect of Early Treatment Group on Calibration Aggregated Across Entire Year

	(1)	(2)	(3)	(4)	(5)
N=2,624	Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	-0.040** (0.010)	0.024* (0.009)	-0.022** (0.006)	-0.041** (0.012)	-0.098** (0.033)
Pretest Accuracy	0.280*** (0.028)	0.050 (0.042)	0.407*** (0.026)	0.786*** (0.055)	1.986*** (0.168)
Grade 4	-0.012 (0.010)	0.015 (0.011)	-0.002 (0.009)	-0.001 (0.019)	0.003 (0.050)
Male	0.044*** (0.008)	-0.057*** (0.008)	0.007 (0.004)	0.006 (0.011)	-0.033 (0.032)
Asian	0.027 (0.017)	-0.031 (0.023)	0.012 (0.009)	0.013 (0.018)	0.001 (0.054)
White	-0.001 (0.013)	0.017 (0.012)	0.013* (0.005)	0.025 (0.012)	0.080* (0.030)
Other Ethnicity	-0.015 (0.013)	0.002 (0.016)	-0.000 (0.009)	-0.004 (0.017)	-0.057 (0.053)
Eng. Lang. Learner	-0.021** (0.007)	0.014 (0.009)	-0.010 (0.005)	-0.013 (0.010)	-0.043 (0.025)
Free/Reduced Lunch	-0.017 (0.010)	0.014 (0.012)	-0.003 (0.006)	-0.008 (0.015)	-0.021 (0.046)
Math CST 2011	0.00004 -0.0001	-0.0001 -0.0001	0.00003 -0.0001	0.00003 -0.0001	-0.0002 -0.0003
ELA CST 2011	0.00002 -0.0001	0.0003** -0.0001	0.0003** -0.0001	0.001** -0.0002	0.002*** -0.0004
Total Objectives	0.046** (0.014)	-0.039* (0.018)	0.015 (0.014)	0.017 (0.032)	0.021 (0.078)
Constant	0.649*** (0.040)	0.319*** (0.042)	0.348*** (0.028)	-0.181*** (0.043)	-0.281 (0.140)
R2	0.135	0.055	0.430	0.313	0.267

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) regression coefficients presented. Standard errors are in parentheses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. The specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 5

Effect of Early Treatment Group on Calibration for Place Value, Robustness Check

		(1)	(2)	(3)	(4)	(5)
N=1,214		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.040***	0.001	-0.035**	-0.047	-0.188*
	SE	(0.010)	(0.015)	(0.011)	(0.025)	(0.083)
	Beta	-0.095***	0.001	-0.078**	-0.045	-0.056*
Pretest Accuracy	B	0.078**	-0.007	0.153***	0.220**	0.544*
	SE	(0.025)	(0.036)	(0.030)	(0.071)	(0.233)
	Beta	0.089**	-0.005	0.163***	0.101**	0.077*
Male	B	0.010	-0.037	-0.005	-0.020	-0.116
	SE	(0.011)	(0.019)	(0.011)	(0.023)	(0.080)
Asian	B	0.040*	-0.001	0.026	0.074	0.223
	SE	(0.016)	(0.047)	(0.018)	(0.044)	(0.218)
White	B	0.002	0.007	0.005	0.022	0.027
	SE	(0.012)	(0.019)	(0.009)	(0.019)	(0.066)
Other Ethnicity	B	-0.002	0.075	0.032	0.074	0.341*
	SE	(0.032)	(0.037)	(0.019)	(0.052)	(0.158)
Eng Lang Learner	B	-0.026	0.012	-0.026*	-0.061*	-0.092
	SE	(0.013)	(0.016)	(0.011)	(0.027)	(0.074)
Free/Reduced Lunch	B	0.001	0.001	-0.007	0.026	0.030
	SE	(0.014)	(0.012)	(0.010)	(0.031)	(0.074)
Math CST 2010	B	0.0003**	0.0002	0.001***	0.001***	0.003**
	SE	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.001)
ELA CST 2010	B	-0.00001	0.0002	0.0002	0.0003	0.001
	SE	(0.0001)	(0.0002)	(0.0002)	(0.0004)	(0.001)
Constant	B	0.776***	0.270***	0.492***	0.019	0.073
	SE	(0.051)	(0.068)	(0.054)	(0.113)	(0.382)
	R2	0.059	0.013	0.155	0.096	0.067

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to fourth graders who had 2010 CST test score data and who had data on the Place Value objective, the first or second objective for all fourth graders. Standard errors clustered on school (N=18).

Table 6

Effect of Early Treatment Group on Calibration for First Three Objectives Encountered in ST Math, Robustness Check

		(1)	(2)	(3)	(4)	(5)
N=1,238		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.058***	0.032**	-0.036**	-0.054*	-0.146*
	SE	(0.011)	(0.010)	(0.011)	(0.021)	(0.061)
	Beta	-0.154***	0.072**	-0.105**	-0.069*	-0.065*
Pretest Accuracy	B	0.175***	0.042	0.291***	0.533***	1.397***
	SE	(0.040)	(0.048)	(0.043)	(0.082)	(0.283)
	Beta	0.172***	0.035	0.319***	0.257***	0.232***
Male	B	0.022*	-0.043**	-0.009	-0.015	-0.091
	SE	(0.010)	(0.011)	(0.007)	(0.016)	(0.049)
Asian	B	0.052*	-0.002	0.034*	0.079	0.305
	SE	(0.021)	(0.039)	(0.014)	(0.041)	(0.151)
White	B	-0.007	0.029**	0.026**	0.049*	0.136
	SE	(0.014)	(0.009)	(0.009)	(0.023)	(0.074)
Other Ethnicity	B	-0.007	0.015	0.007	0.038	0.086
	SE	(0.029)	(0.031)	(0.012)	(0.051)	(0.121)
Eng Lang Learner	B	-0.018	0.013	-0.020	-0.035	-0.059
	SE	(0.009)	(0.014)	(0.011)	(0.024)	(0.070)
Free/Reduced Lunch	B	-0.014	0.008	-0.002	0.012	-0.012
	SE	(0.008)	(0.017)	(0.008)	(0.029)	(0.074)
Math CST 2010	B	0.0003	0.0001	0.001***	0.001***	0.002**
	SE	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.001)
ELA CST 2010	B	-0.00003	0.0002	0.0002	0.0003	0.001
	SE	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.001)
Constant	B	0.320	0.112	0.045	-0.898*	-2.612*
	SE	(0.299)	(0.142)	(0.189)	(0.381)	(1.050)
R2		0.134	0.061	0.344	0.226	0.179

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to fourth graders who had 2010 CST test score data and who had data on at least three objectives, the specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 7

Effect of Early Treatment Group on Calibration Aggregated Across Entire Year, Robustness Check

		(1)	(2)	(3)	(4)	(5)
N=1,238		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.046***	0.041**	-0.014	-0.023	-0.038
	SE	(0.010)	(0.011)	(0.008)	(0.016)	(0.041)
Pretest Accuracy	Beta	-0.125***	0.101**	-0.054	-0.041	-0.025
	B	0.277***	0.060	0.409***	0.808***	1.989***
Male	SE	(0.036)	(0.041)	(0.026)	(0.056)	(0.162)
	Beta	0.215***	0.042	0.445***	0.406***	0.365***
Asian	B	0.044**	-0.061***	-0.002	-0.011	-0.068
	SE	(0.012)	(0.013)	(0.006)	(0.015)	(0.034)
White	B	0.056**	-0.059	0.022	0.027	0.030
	SE	(0.019)	(0.031)	(0.012)	(0.028)	(0.072)
Other Ethnicity	B	0.004	0.023	0.026***	0.055***	0.153***
	SE	(0.014)	(0.015)	(0.006)	(0.013)	(0.035)
Eng Lang Learner	B	-0.044	0.023	-0.007	-0.037	-0.096
	SE	(0.024)	(0.030)	(0.023)	(0.051)	(0.143)
Free/Reduced Lunch	B	-0.025*	0.012	-0.011	-0.016	-0.075
	SE	(0.010)	(0.016)	(0.010)	(0.022)	(0.057)
Math CST 2010	B	-0.015	0.022	0.002	0.008	0.020
	SE	(0.013)	(0.017)	(0.009)	(0.022)	(0.062)
ELA CST 2010	B	0.0001	0.0001	0.0003***	0.001***	0.001**
	SE	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0004)
Total Objectives	B	0.0001	0.0001	0.0001	0.00004	0.0003
	SE	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0004)
Constant	B	0.068	-0.030	0.043	0.080	0.116
	SE	(0.051)	(0.056)	(0.028)	(0.061)	(0.083)
	B	0.550***	0.315***	0.258***	-0.357*	-0.723**
	SE	(0.076)	(0.076)	(0.063)	(0.130)	(0.183)
	R2	0.131	0.063	0.428	0.311	0.268

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to fourth graders who had 2010 CST test score data. The specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 8

Association between Early Treatment Group and Calibration for Place Value, First Year of Treatment

N=2,520		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.036*	-0.016	-0.057***	-0.102***	-0.308***
	SE	(0.014)	(0.009)	(0.011)	(0.024)	(0.067)
	Beta	-0.089*	-0.025	-0.129***	-0.095***	-0.089***
Pretest Accuracy	B	0.134***	0.124**	0.321***	0.542***	1.574***
	SE	(0.015)	(0.033)	(0.022)	(0.042)	(0.117)
	Beta	0.174***	0.102**	0.379***	0.263***	0.237***
Male	B	0.005	-0.041*	-0.011	-0.024	-0.142
	SE	(0.008)	(0.015)	(0.012)	(0.029)	(0.077)
Asian	B	0.021	0.018	0.030	0.064	0.290*
	SE	(0.013)	(0.020)	(0.017)	(0.043)	(0.110)
White	B	0.019**	-0.010	0.010	0.006	0.030
	SE	(0.006)	(0.009)	(0.009)	(0.015)	(0.051)
Other Ethnicity	B	0.017	0.014	0.026	0.034	0.184
	SE	(0.015)	(0.037)	(0.030)	(0.071)	(0.267)
Eng Lang Learner	B	-0.006	-0.017	-0.041***	-0.064**	-0.146*
	SE	(0.007)	(0.010)	(0.009)	(0.020)	(0.055)
Free/Reduced Lunch	B	-0.016	0.012	0.001	-0.011	-0.018
	SE	(0.011)	(0.017)	(0.011)	(0.025)	(0.075)
Constant	B	0.855***	0.329***	0.644***	0.374***	0.946***
	SE	(0.015)	(0.025)	(0.017)	(0.035)	(0.100)
R2		0.045	0.017	0.186	0.089	0.074

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to students who had data on the Place Value objective, the first objective for second and third graders and the first or second objective for fourth graders. Standard errors clustered on school (N=18).

Table 9

Association between Early Treatment Group and Calibration for First Three Objectives, First Year of Treatment

N=2,612		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.035**	-0.044**	-0.094***	-0.172***	-0.449***
	SE	(0.012)	(0.012)	(0.008)	(0.017)	(0.035)
	Beta	-0.098**	-0.095**	-0.276***	-0.220***	-0.195***
Pretest Accuracy	B	0.215***	0.125***	0.419***	0.756***	2.021***
	SE	(0.020)	(0.023)	(0.021)	(0.032)	(0.079)
	Beta	0.235***	0.108***	0.483***	0.381***	0.346***
Male	B	0.022*	-0.043***	-0.003	-0.010	-0.082*
	SE	(0.008)	(0.009)	(0.006)	(0.014)	(0.038)
Asian	B	0.001	0.004	0.018	0.027	0.055
	SE	(0.017)	(0.018)	(0.010)	(0.028)	(0.083)
White	B	0.018*	0.018	0.027**	0.058***	0.192***
	SE	(0.007)	(0.009)	(0.008)	(0.014)	(0.042)
Other Ethnicity	B	-0.003	0.041	0.012	0.047	0.188
	SE	(0.019)	(0.022)	(0.017)	(0.035)	(0.146)
Eng Lang Learner	B	-0.010	-0.003	-0.029***	-0.046**	-0.088*
	SE	(0.007)	(0.008)	(0.007)	(0.013)	(0.036)
Free/Reduced Lunch	B	0.002	-0.012	-0.000	-0.014	-0.049
	SE	(0.009)	(0.015)	(0.010)	(0.024)	(0.075)
Constant	B	0.824***	0.087	0.336**	-0.172	-0.559
	SE	(0.099)	(0.133)	(0.098)	(0.204)	(0.537)
R2		0.094	0.048	0.376	0.237	0.195

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Sample limited to students who had data on at least three objectives, the specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 10

Association between Early Treatment Group and Calibration Aggregated Across All Objectives, First Year of Treatment

N=2,612		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.034*	0.007	-0.035***	-0.064***	-0.157**
	SE	(0.013)	(0.016)	(0.006)	(0.012)	(0.040)
	Beta	-0.102*	0.018	-0.132***	-0.114***	-0.100**
Pretest Accuracy	B	0.301***	0.106*	0.513***	0.978***	2.457***
	SE	(0.028)	(0.042)	(0.021)	(0.042)	(0.117)
	Beta	0.272***	0.084*	0.589***	0.532***	0.478***
Male	B	0.036***	-0.057***	-0.002	-0.016	-0.079**
	SE	(0.008)	(0.008)	(0.004)	(0.010)	(0.021)
Asian	B	0.007	-0.031	-0.006	-0.021	-0.101
	SE	(0.019)	(0.015)	(0.016)	(0.033)	(0.085)
White	B	0.009	0.003	0.008	0.015	0.058
	SE	(0.008)	(0.010)	(0.006)	(0.011)	(0.032)
Other Ethnicity	B	-0.010	0.023	0.008	0.020	0.065
	SE	(0.014)	(0.015)	(0.010)	(0.018)	(0.063)
Eng Lang Learner	B	-0.014*	0.002	-0.015***	-0.025**	-0.075**
	SE	(0.006)	(0.008)	(0.003)	(0.008)	(0.022)
Free/Reduced Lunch	B	0.001	-0.005	-0.004	-0.006	-0.024
	SE	(0.008)	(0.011)	(0.005)	(0.012)	(0.035)
Total Objectives	B	0.008**	-0.014***	-0.006*	-0.009	-0.025
	SE	(0.003)	(0.003)	(0.002)	(0.005)	(0.018)
Constant	B	0.680***	0.319***	0.399***	-0.096	-0.090
	SE	(0.039)	(0.053)	(0.028)	(0.066)	(0.241)
R2		0.107	0.050	0.479	0.366	0.305

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Calibration and pretest accuracy aggregated across the entire year's curriculum; the specific objectives included were controlled as a series of dummy variables (omitted). Standard errors clustered on school (N=18).

Table 11

Association between Early Treatment Group and Calibration Aggregated Across All Objectives, First Year of Treatment, Limited to Fourth Graders in 2011

N=1,259		(1)	(2)	(3)	(4)	(5)
		Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Early Treat Group	B	-0.028	0.015	-0.020	-0.039	-0.058
	SE	(0.020)	(0.026)	(0.015)	(0.030)	(0.098)
	Beta	-0.084	0.039	-0.076	-0.070	-0.037
Pretest Accuracy	B	0.315***	0.034	0.468***	0.906***	2.139***
	SE	(0.050)	(0.058)	(0.028)	(0.057)	(0.149)
	Beta	0.279***	0.026	0.527***	0.473***	0.404***
Male	B	0.039**	-0.063***	-0.008	-0.029*	-0.103**
	SE	(0.012)	(0.013)	(0.005)	(0.012)	(0.032)
Asian	B	0.048**	-0.070**	0.017	0.017	-0.042
	SE	(0.016)	(0.022)	(0.012)	(0.028)	(0.065)
White	B	0.022*	0.006	0.020**	0.041**	0.144**
	SE	(0.009)	(0.016)	(0.006)	(0.013)	(0.043)
Other Ethnicity	B	-0.027	0.027	0.001	-0.004	-0.003
	SE	(0.022)	(0.028)	(0.015)	(0.023)	(0.077)
Eng Lang Learner	B	-0.012	0.009	-0.012	-0.015	-0.038
	SE	(0.013)	(0.014)	(0.009)	(0.022)	(0.059)
Free/Reduced Lunch	B	-0.005	0.011	-0.002	0.004	0.015
	SE	(0.016)	(0.017)	(0.009)	(0.021)	(0.055)
Math CST 2010	B	-0.0001	0.0001	0.0002*	0.0003	0.001
	SE	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0003)
ELA CST 2010	B	0.00001	0.0003	0.0001	0.0002	0.001*
	SE	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0005)
Total Objectives	B	0.008	-0.019*	-0.004	-0.006	-0.038
	SE	(0.006)	(0.007)	(0.005)	(0.010)	(0.031)
Constant	B	0.604***	0.236***	0.339***	-0.220**	-0.796***
	SE	(0.041)	(0.056)	(0.031)	(0.059)	(0.178)
R2		0.126	0.076	0.503	0.377	0.324

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. The reference group comprises students who were females, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Calibration and pretest accuracy aggregated across the entire year's curriculum; the specific objectives included were controlled as a series of dummy variables (omitted). Sample limited to those students who were in fourth grade in 2011 and had data on 2010 CST scores. Standard errors clustered on school (N=18).

Table 12

Correlations between Objectives, Pretest Accuracy, Third Grade (Bottom) & Fourth Grade (Top)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	0.160	0.102 ^c	0.098 ^c	0.120 ^c	0.190	0.063 ^a	0.052 ^a	0.188	0.149	0.061 ^a	0.143	0.211	0.237	0.227	0.094 ^b	0.089 ^b
2	0.350	1	0.325	0.395	0.188	0.219	0.123	0.240	0.294	0.253	0.241	0.108 ^c	0.243	0.225	0.213	0.146	0.174
3	0.309	0.315	1	0.416	0.191	0.224	0.131	0.222	0.259	0.264	0.242	0.056 ^a	0.235	0.234	0.228	0.174	0.166
4	0.226	0.165	0.113	1	0.267	0.304	0.211	0.280	0.364	0.355	0.357	0.096 ^c	0.362	0.265	0.269	0.201	0.222
5	0.220	0.245	0.212	0.201	1	0.178	0.175	0.144	0.175	0.224	0.184	0.072 ^b	0.232	0.141	0.128	0.100 ^b	0.199
6	0.184	0.158	0.166	0.150	0.216	1	0.240	0.241	0.277	0.321	0.241	0.113	0.311	0.238	0.290	0.16	0.213
7	0.307	0.333	0.325	0.151	0.151	0.189	1	0.203	0.175	0.201	0.212	0.146	0.124	0.163	0.127	0.055 ^a	0.248
8	0.248	0.238	0.223	0.209	0.197	0.191	0.321	1	0.326	0.315	0.319	0.063 ^a	0.265	0.226	0.272	0.173	0.278
9	0.278	0.301	0.291	0.209	0.221	0.273	0.304	0.312	1	0.331	0.287	0.076 ^b	0.313	0.277	0.218	0.173	0.238
10	0.230	0.307	0.262	0.134	0.255	0.219	0.279	0.279	0.302	1	0.361	0.102 ^c	0.330	0.333	0.332	0.239	0.230
11	0.171	0.194	0.218	0.237	0.201	0.208	0.161	0.272	0.253	0.217	1	0.106 ^c	0.369	0.271	0.301	0.226	0.316
12	0.210	0.206	0.207	0.268	0.219	0.195	0.214	0.282	0.258	0.263	0.287	1	0.115	0.122 ^c	0.088 ^b	0.126 ^c	0.114 ^c
13	0.112 ^c	0.098 ^c	0.173	0.155	0.109 ^c	0.159	0.142	0.212	0.151	0.170	0.256	0.156	1	0.3	0.298	0.27	0.270
14	0.248	0.349	0.323	0.089 ^b	0.236	0.233	0.327	0.257	0.296	0.297	0.179	0.241	0.117 ^c	1	0.350	0.200	0.266
15	0.242	0.294	0.260	0.143	0.241	0.194	0.203	0.204	0.326	0.277	0.259	0.283	0.136	0.318	1	0.284	0.306
16	0.175	0.201	0.138	0.242	0.239	0.161	0.196	0.189	0.200	0.224	0.206	0.190	0.155	0.145	0.223	1	0.213
17	0.219	0.322	0.260	0.169	0.177	0.256	0.275	0.205	0.306	0.260	0.193	0.327	0.176	0.314	0.311	0.200	1
18	0.184	0.252	0.320	0.109 ^c	0.180	0.103 ^c	0.332	0.249	0.261	0.283	0.201	0.177	0.142	0.267	0.269	0.122 ^c	0.212

Note. All correlations at $p < .001$ level except ^a $p > .05$, ^b $p < .05$, ^c $p < .01$. Correlations of quiz accuracy at pretest between objectives.

Third grade shown below the diagonal, fourth grade above. Objectives limited to those that at least 50% of the students completed across both the ETG and LTG.

Table 13

Association between Calibration Gain and Posttest Performance Gain, Paired Quizzes

N=1,586	(1)	(2)	(3)	(4)	(5)
	Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Gain	0.048* (0.018)	-0.041** (0.011)	-0.032 (0.027)	0.0001 (0.008)	-0.001 (0.003)
Pretest Gain	0.222*** (0.031)	0.237*** (0.030)	0.248*** (0.035)	0.232*** (0.033)	0.234*** (0.033)
Early Treatment Group	0.032* (0.012)	0.031* (0.011)	0.035** (0.012)	0.034** (0.011)	0.034** (0.011)
Grade 4	-0.085** (0.023)	-0.089** (0.023)	-0.094** (0.025)	-0.092** (0.024)	-0.093** (0.024)
Male	-0.008 (0.010)	-0.008 (0.010)	-0.007 (0.010)	-0.007 (0.010)	-0.007 (0.010)
Asian	0.028 (0.025)	0.031 (0.026)	0.032 (0.025)	0.032 (0.025)	0.032 (0.025)
White	0.005 (0.009)	0.005 (0.009)	0.006 (0.009)	0.006 (0.009)	0.006 (0.009)
Other Ethnic	-0.016 (0.025)	-0.020 (0.024)	-0.018 (0.025)	-0.018 (0.025)	-0.018 (0.025)
English Lang. Learner	0.002 (0.010)	0.001 (0.010)	0.001 (0.011)	0.001 (0.010)	0.001 (0.010)
Free/Reduced Lunch	-0.001 (0.011)	-0.001 (0.012)	-0.001 (0.011)	-0.001 (0.011)	-0.001 (0.011)
Constant	0.012 (0.018)	0.014 (0.018)	0.013 (0.019)	0.013 (0.019)	0.013 (0.019)
R2	0.168	0.169	0.165	0.164	0.164

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients presented. Standard errors are in parentheses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). Sample limited to those who had data on the two selected quizzes for this analysis.

Table 14
Association between Calibration Gain and Math CST Gain

	(1)	(2)	(4) Simple Match	(5) Gamma	(6) Discrim.
N=1,586	Sensitivity	Specificity			
Calibration Gain	-8.186 (5.545)	5.639 (4.205)	2.528 (4.331)	-2.314 (2.392)	-0.170 (0.637)
Early Treatment Group	-1.155 (5.682)	-1.256 (5.693)	-1.740 (5.570)	-1.484 (5.651)	-1.600 (5.656)
Grade 4	-25.219* (8.994)	-23.687* (9.345)	-22.954* (9.570)	-24.420* (9.487)	-23.602* (9.631)
Male	2.318 (3.269)	2.330 (3.249)	2.211 (3.240)	2.161 (3.234)	2.208 (3.244)
Asian	-9.334 (8.094)	-9.747 (8.241)	-9.928 (8.199)	-9.765 (8.084)	-9.891 (8.138)
White	-2.575 (6.535)	-2.609 (6.498)	-2.640 (6.532)	-2.661 (6.554)	-2.648 (6.520)
Other Ethnic	11.697 (10.900)	12.225 (10.844)	12.015 (10.740)	12.000 (10.684)	11.958 (10.764)
English Lang .Learner	15.291** (4.002)	15.446** (3.955)	15.476** (3.958)	15.535** (3.969)	15.498** (3.967)
Free/Reduced Lunch	-0.379 (5.346)	-0.437 (5.401)	-0.382 (5.420)	-0.487 (5.375)	-0.428 (5.404)
Constant	17.600 (8.918)	17.046 (9.094)	17.077 (9.116)	17.680 (8.981)	17.347 (9.077)
R2	0.054	0.054	0.052	0.053	0.052

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients presented. Standard errors are in parentheses. The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). Sample limited to those who had data on the two selected quizzes for this analysis.

Table 15

Association between Calibration Growth and Growth of Quiz Posttest Performance

	(1)	(2)	(3)	(4)	(5)
N=2,625	Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Slope	0.036 (0.117)	0.048 (0.120)	0.183 (0.122)	0.062 (0.044)	0.023 (0.018)
Pretest Acc.	-0.016 (0.009)	-0.017 (0.010)	-0.017 (0.009)	-0.017 (0.009)	-0.017 (0.010)
Early Treatment Group	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)
Grade 4	-0.011** (0.003)	-0.011** (0.003)	-0.011** (0.003)	-0.011** (0.004)	-0.011** (0.004)
Male	0.000 (0.001)	0.000 (0.001)	0.001 (0.001)	0.001 (0.001)	0.000 (0.001)
Asian	-0.004* (0.002)	-0.004* (0.002)	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.002)
White	-0.002 (0.003)	-0.002 (0.003)	-0.001 (0.003)	-0.002 (0.003)	-0.002 (0.003)
Other Ethnic	-0.002 (0.002)	-0.002 (0.002)	-0.002 (0.002)	-0.002 (0.002)	-0.002 (0.002)
English Lang. Learner	0.000 (0.001)	0.000 (0.001)	-0.000 (0.002)	0.000 (0.001)	-0.000 (0.001)
Free/Reduced Lunch	-0.004* (0.002)	-0.004* (0.002)	-0.004 (0.002)	-0.004* (0.002)	-0.004* (0.002)
Total Objectives	0.007 (0.019)	0.006 (0.019)	0.007 (0.018)	0.006 (0.018)	0.006 (0.018)
Constant	0.011 (0.019)	0.012 (0.019)	0.009 (0.018)	0.009 (0.019)	0.011 (0.019)
R2	0.069	0.071	0.092	0.090	0.089

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients presented. Standard errors are in parentheses. Specific objectives tested controlled with a series of dummy variables (not shown). The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18).

Table 15

Association between Calibration Growth and End-of-Year Math CST Scores

	(1)	(2)	(3)	(4)	(5)
N=2,624	Sensitivity	Specificity	Simple Match	Gamma	Discrim.
Calibration Slope	-0.901 (10.441)	17.112 (9.828)	48.919* (18.402)	9.754 (6.984)	3.063 (2.599)
Pretest Acc.	100.489*** (10.970)	100.242*** (10.920)	100.223*** (10.751)	100.402*** (10.919)	100.393*** (10.873)
Early Treatment Group	1.857 (2.819)	1.872 (2.838)	1.823 (2.821)	1.783 (2.834)	1.802 (2.838)
Grade 4	11.840 (7.373)	11.768 (7.385)	11.833 (7.292)	11.823 (7.354)	11.820 (7.372)
Male	4.853** (1.592)	4.856** (1.584)	4.999** (1.571)	4.902** (1.597)	4.863** (1.586)
Asian	8.218 (4.536)	8.243 (4.541)	8.419 (4.515)	8.277 (4.532)	8.270 (4.514)
White	0.635 (2.990)	0.636 (3.011)	0.847 (3.006)	0.683 (3.014)	0.683 (3.001)
Other Ethnic	5.155 (7.603)	5.227 (7.645)	5.235 (7.588)	5.278 (7.651)	5.274 (7.637)
English Lang Learner	-2.260 (2.790)	-2.303 (2.796)	-2.333 (2.783)	-2.247 (2.772)	-2.275 (2.779)
Free/Reduced Lunch	-2.663 (2.188)	-2.624 (2.182)	-2.476 (2.175)	-2.632 (2.200)	-2.599 (2.175)
Total Objectives	-10.101 (6.615)	-10.270 (6.639)	-9.868 (7.093)	-10.054 (6.826)	-10.120 (6.791)
Math CST 2011	0.258*** (0.029)	0.258*** (0.029)	0.258*** (0.029)	0.258*** (0.029)	0.258*** (0.029)
ELA CST 2011	0.230*** (0.039)	0.231*** (0.039)	0.232*** (0.039)	0.231*** (0.039)	0.231*** (0.039)
Constant	105.763*** (17.527)	105.613*** (17.459)	104.637*** (17.189)	105.184*** (17.488)	105.504*** (17.486)
R2	0.643	0.643	0.643	0.643	0.643

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Unstandardized regression coefficients presented. Standard errors are in parentheses. Specific objectives included controlled with a series of dummy variables (not shown). The reference group comprises students who were females in third grade, Hispanic, Non-ELL, and not eligible for free/reduced lunch. Standard errors clustered on school (N=18). One student omitted who did not have ELA CST data for 2011.