A Math Intervention Model for Middle School: How the Combination of Formative Assessment, Feedback, Academic Vocabulary, and Word Problems Affect Student Achievement in Mathematics

By:
Heather Karuza

Claremont Graduate University
2014

A dissertation submitted to the faculty of Claremont Graduate University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Education

Committee Chair: Mary Poplin, Ph.D.
Committee Member: Carl Cohn, Ed.D.
Committee Member: June Hilton, Ph.D.

APPROVAL OF THE REVIEW COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Heather Karuza as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy of Education.

Dr. Mary Poplin, Chair
Claremont Graduate University
School of Educational Studies

Dr. Carl Cohn
Claremont Graduate University
School of Educational Studies

Dr. June Hilton
Claremont Graduate University
School of Educational Studies

Abstract of the Dissertation

A Math Intervention Model for Middle School: How the Combination of Formative Assessment, Feedback, Academic Vocabulary, and Word Problems Affect Student Achievement in Mathematics

By:
Heather Karuza

Claremont Graduate University: 2014

This study determines the effect of a math intervention model that combines the use of formative assessment, feedback, academic vocabulary, and word problems on student achievement scores in an urban middle school ($N \approx 1000$ each year) from 2010 to 2012. The researcher developed the DARTS model which is an acronym that stands for: Diagnostic, Assessment, Rescue Assignment, Translations, and Story Problems. The model is largely based on the theory of feedback by Hattie and Timperley (2007), and was implemented school-wide in the treatment school beginning in the 2009-2010 school year. The data used for this study includes California Standards Test (CST) math scaled scores, Academic Growth over Time (AGT) school-wide scores, and band-level percentages. T-tests were run to look for significant differences between the years before and during the implementation of the DARTS model at the treatment school. In addition, t-tests were run on the treatment school's CST English, Science, and History scores to compare them to the changes in math scores. The treatment school scores were also compared to scores from two non-treatment similar schools.

The results showed that the mean CST math scaled scores increased significantly for all three grade levels at the treatment school in 2010 and 2011. This was not the case, however, for the other content areas (English, Science, and History). The treatment school's math scores were significantly higher than both similar schools during the years that DARTS was implemented. Seven out of seven t-tests showed the treatment school significantly out-performed the respective

similar school from 2010 to 2012. AGT scores were higher for the treatment school than the similar schools in 8 out of 10 comparisons for math, but only in 3 out of 14 for other content area comparisons.

Overall, the DARTS model was effective in increasing math student achievement scores in the treatment school. Although this study has some limitations, it can be useful for schools looking for a research-based, comprehensive, and feasible intervention model for mathematics that can help raise student achievement.

Dedication

My work on this dissertation is dedicated to the following people:

Coleen Kaiwi, my former Principal, mentor, and professional role model. She took a chance on me, bringing me out of the classroom to coach other teachers. She believed I could make a greater impact on students through my leadership. Her mentorship helped develop me as a leader and her support ensured that the DARTS model could be implemented with fidelity. Ms. Kaiwi always had my back and pushed me to do better.

Reginald Brookens, who helped me develop the DARTS model and make it successful in the treatment school. He believed I had what it took to run a program like this and brought other schools over to see it in action.

The math teachers and students at the treatment school who embraced the DARTS model and implemented it to the best of their abilities. After taking an honest look at the data, the teachers agreed that we had to do things differently. They “bought-in” to the idea of using data to drive instruction and did all of the components of the DARTS model on top of an already demanding job. They are the reason that the DARTS model was truly successful.

Acknowledgements

This entire study would not have been possible without the guidance and encouragement from my mentor, Dr. Mary Poplin. She believed in me and this study from the first day we met. When she agreed I could do my dissertation on my DARTS model, I was elated. Without her support, I would not have been able to follow my passion, and it truly means so much to me. I am forever grateful for my committee members, Dr. Carl Cohn and Dr. June Hilton. Dr. Cohn created a cohort to meet at CSU Dominguez Hills, and without that connection, I may have never pursued this degree. His love and support of our cohort has helped us make large impacts in urban education. I especially thank Dr. Hilton for meeting with me so many times to help with my statistics and stepping in when I really needed her.

I want to thank my Principal, Pedro Garcia, for his unwavering support of not only the DARTS model, but of my Ph.D. journey and leadership. He truly believes in me, and his continuous support and mentorship have helped me become a better leader each day. I would also like to thank Dr. Ann Chlebicki, Dr. Antonia Issa Lahera, and Dr. Anthony Normore for seeing something special in me and giving me a push to get my doctoral degree. I continue to be inspired by their dedication to urban school leadership.

Thank you to my cohort members: Ruth, Mona, Nancy, Kim, Candyce, Crystal, Ruben, Eugene, Gabrielle, Cassandra, Brian, Javier, Lanette and Ashley. I would also like to thank Dr. Lynn Winters for her mentorship along the way and for always being there for us. Our cohort had a blast getting to know each other, and throughout this stressful ride, we have become life-long friends and colleagues. Thank you all for your support and friendship! I have learned so much from all of you.

Finally, I want to acknowledge my family for their encouragement and love. My Mom and Dad who have loved me unconditionally, taught me the importance of education, and never accepted anything less than my best. Most of all, I want to thank and acknowledge my husband, Jason Karuza. I could not have done this without his undying love, support, and sacrifice! He has always been my #1 fan, and now that this is done, I owe him 3 years-worth of fun-filled weekends!

Table of Contents

List of Tables

Chapter I: Introduction

Overview

Math achievement scores have declined in each successive year of the middle grades (6th through 8th) in the state of California. These scores decline even more for students located in urban areas due to numerous additional challenges. By the time students reach middle school, large numbers of them enter with gaps in their math skills and knowledge. Teachers struggle to integrate strategies that will help students catch up to grade level while they continue to teach the core content of their class. Even when students are programmed into a math intervention course, in many schools there is no set curriculum for teachers to follow and no concrete way for them to collect necessary data for assessing the students' needs. There is a need for a successful math intervention model that is both research-based and affordable for schools.

Purpose of the Study

The purpose of this study is to understand how the implementation of the DARTS math intervention model affected math achievement scores of approximately 1000 urban students (each year) in 6th grade, 7th grade, and 8th grade Algebra 1 in a large urban middle school between the years of 2009 to 2012. DARTS is a comprehensive model that combines research-based strategies in order to make math intervention feasible and successful. This study will analyze how students did over time from 2007 to 2012 in the treatment school, while also comparing students at the treatment school to students at similar schools who did not implement the model.

Significance of the Study

This study is important for three main reasons. First, California Standards Test (CST) math scores in urban districts decline beginning with the transition from 5th to 6th grade and continue that negative trend through high school. Data from the CST Algebra 1 scores show that, on average, 64% of students in the state of California and 70% of the students in the state's largest district did not meet the proficiency level in 2012. Average Algebra 1 scaled scores for the Los Angeles Unified School District (LAUSD), which is the largest urban district in California, are below the state average, and the district continues to focus on improving these numbers (Los Angeles Unified School District, 2012). The mean scaled score for 8th graders in Algebra 1 that same year was 355.1 in California, but only 329.7 in LAUSD. These data do not include the over 40% of 8th graders who were not enrolled in Algebra 1 (California Department of Education, 2013). These statistics point to an issue of preparedness. Many students come into middle school with a lack of basic mathematical skills and therefore struggle or fail to meet the more rigorous 6th and 7th grade standards, which means they will not be ready for Algebra. There are tremendous gaps in their skills, which can prevent them from accessing higher-level math. Indeed, the Introduction to the Common Core Standards for Mathematics state that "what students can learn at any particular grade level depends upon what they have learned before" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010, p. 5). In addition, the aforementioned introduction also mentions that there are no definite intervention methods or materials to help students at this time.

There is some research available regarding the transitions from elementary to middle and high school (Barber & Olsen, 2004), and the typical negative impact those changes have on test scores (Alspaugh, 1998). However, little research exists on why the math scores continuously drop between the grade levels within middle school, and what models are successful in alleviating this decline, especially in urban areas.

Success in middle school mathematics is critical for student success in high school mathematics and high school graduation (Silver, Saunders, & Zarate, 2008; Smith Green, 2012; Wang & Goldschmidt, 2003), yet little emphasis is placed on improving math instruction and intervention in the middle grades. In the research by Silver et al. (2008), which is an analysis on the factors leading to dropouts in LAUSD, students who did not graduate on time failed an average of 3.5 times as many middle school courses as those who did graduate. Students who failed classes in 8^{th} grade were at an even greater risk of dropping out of high school, and “less than 20% of students who failed classes in all three years of middle school graduated from high school” (Silver et al., 2008, p. 14). Even more important is their finding of the failure rate in Algebra 1: 65% of students fail each year on average, and students who do not pass Algebra 1 by the 9^{th} grade are half as likely to graduate from high school (35% vs. 70%).

Second, there are many research-based strategies that can help teachers improve their instruction, but strategies only work well when there is quality curriculum with which to utilize them. There are few comprehensive models that can be employed at any school, on any level, and on top of any curriculum. Any models that do exist are expensive, require students to use technology that may not be readily available, or are not feasible for implementation. Teachers are expected to differentiate instruction to meet the needs of each student, but given no systematic approach to diagnose learning gaps, skill deficits, or misconceptions. They need a way to collect meaningful data in order to provide individualized intervention and build on what students already know.

Third, if the results of this study are as hypothesized, then this model is a viable option for schools as the main curriculum in math intervention classes to increase student achievement. All data, findings, materials, and resources from this study and the DARTS model will be shared with the school district in this study at no cost. Schools will also have the option of requesting professional development on how to successfully implement this model.

Theoretical Foundation

The theory on formative assessments by Black and Wiliam (2009), incorporates five key strategies that focus on the teacher, the peer, and the learner – how they use the information to make decisions, help each other, and help themselves move forward in the learning. The diagnostics provide teachers with specific data to understand what students know, do not know, or have misconceived. The learning theory behind formative assessments is based on feedback loops so that both the student and the teacher have information about what the student does and does not understand. The use of multiple choice answers helps teachers see trends in common mistakes and misconceptions with particular skills. The diagnostic and assessment components utilize frequently collected formative data. Decisions made on what to reteach, review, or reinforce are based on data.

The rescue assignment provides specific ways students can fix their mistakes from the diagnostic, which is based on the feedback theory of Hattie and Timperley (2007). Their work concludes that effective feedback answers three major questions on what the goals are, what progress is being made, and what needs to be done to get to those goals. The rescue assignment enables students to find answers to all of these questions on an individual basis, meaning each student has his/her own targeted outcomes to practice each week. The rescue assignment and the diagnostic work together to form a feedback loop; after working on specific rescue assignment questions, students have the chance to re-assess themselves on the next week's diagnostic which has many of the same problems that the majority of students got incorrect.

Rubenstein and Thompson (2001) emphasize the importance of building mathematics vocabulary and the language of math, which is the theory behind the use of translations. Students have to practice translating mathematical phrases from English into math symbols in order to understand the vocabulary, learn how math is spoken/read, and understand concepts. It is this cognitive understanding of the principles of mathematics that is the theory behind translations; building the use of vocabulary in context.

Finally, the real-world application of mathematics is what is considered "doing mathematics." Mathematics was invented in order to make sense of the world and enable people to speak specifically and quantitatively about the world around them. There is often an emphasis on arithmetic and algorithms in math classes, but students struggle to use these skills in connection to real-life applications. Story problems use a task-specific version of the four-fold graphic organizer based on the visual tools theory by Hyerle (1996). Word problems are difficult for both students and teachers. Having a set way to problem solve helps students determine what the problem is asking, use mathematical strategies to find an answer, and express their solution in multiple ways.

Background

The Trends in International Mathematics and Science Study (TIMSS) brought mathematics education to the forefront by showing that students in the US perform significantly worse in mathematics than students in other countries (Mullis, Martin, Gonzalez, & Chrostowski, 2004). There are many theories as to why students struggle with math and what the essential elements might be in order to alleviate the problem. Of utmost importance is the need to ensure students are prepared for Algebra, and a critical time to begin laying the groundwork for that mastery of the fundamental mathematics begins in 6th grade (Phelan, Choi, Vendlinski, Baker, & Herman, 2011).

In order to improve student achievement, the issue of effective math intervention models must be addressed, and not solely those focused on students with disabilities. According to What Works Clearinghouse (WWC), “only 5 of the more than 40 middle school math interventions known to be available for adoption have any studies of their effectiveness that meet the WWC evidence standards” (What Works Clearinghouse, 2004, p. 2). Those studies that made the cut are costly and require technology to implement. In 2007, WWC identified 158 empirical studies of 34 middle school math programs, only 4 of which met evidence standards, and only 2 of those showed positive results (What Works Clearinghouse, 2007).

Cohen and Hill (2000) studied mathematics reform in California and found that models that incorporate assessment, curriculum, and professional development can have a significant influence on student performance. Their study focused on implications for policy, which could impact practice and therefore affect student achievement. They point out that the California Math Framework for K-12 instruction places an emphasis on mathematical understanding and diminishes the need to memorize math facts. This is an important implication for teachers in the middle and high school grades because they must continue with deeper understandings of grade-level content even if their students lack basic computational skills like adding, subtracting, and multiplication facts. Students who are not native English speakers are also at a disadvantage, since they are also learning the language of math.

It is troubling to see a child struggle with prime factorization in middle school because he does not know his multiplication tables. The researcher developed the D, R, T, and S components of the DARTS math model over the summer of 2009 in order to begin implementation in all three grade levels in the 2009-2010 school year. She ran the model for three years and also created the Google Spreadsheet for data collection/sharing as well as training components and an implementation rubric. The DARTS model works as an intervention to fill in these gaps in math skills for students so that they are better able to access grade-level standards.

The treatment school in this study had dismal math scores in 2009. It was the third year of a negative trend in Algebra 1 scores for 8th grade, plummeting down to only 4% of students reaching proficiency. The 6th and 7th grade scores were not much better at only 20% and 12% proficient respectively. Of the 1,870 students tested that year, 1,197 of them were not proficient in their grade level for math, as measured by the California Standards Test (CST). After the first year of the school implementing the DARTS model, proficiency scores went up to 33% in 6th grade (a 13% increase), 23% in 7th grade (an 11% increase), and 12% in Algebra 1 (an 8% increase). Scores for 7th grade doubled, and Algebra 1 increased 28% after the second year of implementation (Table 1) (scores from http://star.cde.ca.gov).

Table 1: Student Math Proficiency Percentages by Year

	2009	2010	2011
6th Grade	20%	33%	N/A
7th Grade	12%	23%	46%
Algebra 1	4%	12%	40%

Research Questions

This study attempts to determine if the middle-grade students performed better on the CST after having instruction via the DARTS model in their math or math-intervention course. Specifically, the questions for this research are as follows:

1) How did the DARTS math intervention model affect math achievement scores for students in an urban middle school from 2010 to 2012 in grades 6 to 8?
2) To what extent can outside factors (e.g. teachers, school leadership, school-wide initiatives, district reform) be excluded as reasons for changes in the treatment school's student achievement in mathematics during those years?
3) What is the difference between math achievement scores of students in the treatment school who received the DARTS math intervention model compared to those from similar schools who did not implement the model?
4) Does the DARTS model have a greater impact on students from the treatment school who score in the low range (Below Basic and Far Below Basic), middle range (Basic), and/or high range (Proficient and Advanced) of the math CST?

Definition of Terms

DARTS is an acronym that stands for: (D) diagnostic, (A) assessment, (R) rescue assignment, (T) translations, and (S) story problems. These are the components of the DARTS model.

The *diagnostic* is a 10-question multiple-choice (a-d) quiz given once every week in the math class. *Assessment* in the DARTS model differs from the diagnostic in that it is a quarterly static benchmark test used to measure student knowledge of the core course content.

The *rescue assignment* is differentiated intervention for students based on what each student needs to practice. For each of the ten questions on the diagnostic, there are 3-4 free-response questions on the rescue. Each student does the corresponding problems for each problem that s/he missed on the diagnostic. *Translations* are math phrases completely written out in English words, which students translate into math symbols. These provide opportunities to engage in rich mathematical discussions utilizing proper vocabulary. A *story problem* is a mathematical word problem.

In this study the reference to *intervention model* refers to curriculum and instructional strategies that serve to catch students up to grade-level so that they can access and be successful with the core curriculum. It fills in the gaps in student knowledge. The term *urban school* refers to an inner-city school with the majority of the population consisting of minority children from low socioeconomic backgrounds.

Chapter II: Literature Review

Overview

The purpose of this chapter is to review the literature and research regarding the major components of the DARTS Math Intervention Model used for this study. Those components include: formative assessments and formative feedback in educational settings, academic vocabulary in mathematics, and word problems as part of the curriculum. All of these components directly align with those from the DARTS model: diagnostic, assessment, rescue, translations, and story problems.

First, formative assessments are differentiated from summative and other types of assessments. Researchers differ in their perspectives regarding whether formative assessment data are primarily for the teachers' use to inform instruction, or for students to use as feedback for improvement. Next, the frequency of formative assessments is discussed to determine how often teachers should collect formative data. How teachers collect and analyze data are important aspects of how formative assessment data are used to improve student achievement. In order to use these data appropriately, teachers need to have sufficient professional development and training. Coupled closely with formative assessment data is formative feedback. There are various frameworks for feedback, but no matter which one is used, all of the research describes the need for feedback with any formative assessment system. The research on formative assessments is clear about the benefits to student achievement. What exactly constitutes formative assessment is left out because there is an art to developing types that work in various contexts and settings.

Following formative feedback is a review of the research on academic vocabulary and the language of mathematics. A language in its own right, mathematics is a complicated entwinement of symbols and words, and due focus is deserved unto the vocabulary itself. More than just the content vocabulary, this review looks into the academic vocabulary that necessitates the teaching and learning of math within the English language. Teachers must pay attention to the nuances of how mathematics is spoken.

The last part of this literature review centers on the use of word problems in teaching mathematics. Specifically, how graphic organizers and rubrics enhance the understanding and completion of word problems. There are many types of graphic organizers, but overall research in their usage indicates that they have a positive impact in learning outcomes. Rubrics are useful when there is some ambiguity in the grading or performance measure of a particular task or assessment. They provide specific guidance for students if used at the beginning of a task and allow for specific feedback for improvement. This review concludes with implications and recommendations for further research.

Formative and Summative Assessments

In much of the technical literature on assessments, the terms formative and summative are loosely (if at all) defined (Wiliam & Black, 1996). The terms were first given the definition that we use today by Bloom, Hastings, and Madaus (1971). Summative assessments provide evidence of student achievement in order to measure outcomes, assign grades, or evaluate progress. While these types of assessments can arguably be used to inform instruction in some cases, it is most useful to assign the term summative in reference to the function it serves, rather than to the assessment itself (Wiliam & Black, 1996).

Perie, Marion, and Gong (2009) said that data gathered from assessments are used to serve three different purposes: to make instructional, evaluative, and predictive decisions. The term 'formative assessment' is used to distinguish data gathered from a particular assessment and used for instructional decisions rather than for evaluative or predictive decisions. The Association for Supervision and Curriculum Development (ASCD) defines a formative assessment as "an assessment which provides feedback to the teacher for the purpose of improving instruction" (Association for Supervision and Curriculum Development (ASCD), 1996, p. 59). On the other hand, Black and Wiliam (2006) argue that the ultimate user of formative assessment data is the student/learner. It is difficult to argue for which party – the student or the educator – formative assessments are best used. Researchers advocate for the use of data from formative assessments on both sides. In her article, Brookhart (2007) gives the following definition: "formative classroom assessment gives teachers information for instructional decisions and gives pupils information for improvement" (p. 43). Alternatively, formative assessment can be thought of as a process instead of as a thing that teachers give to students. "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (Popham, 2008, p. 5). Whether formative assessment is better used by the teacher or the student, researchers agree that when used correctly, there can be positive results in student achievement. Either way, the data gathered from formative assessments are critical to the "instruction-assessment feedback loop" (Conderman & Hedin, 2012, p. 163) because they guide decisions made by teachers and provide students with an understanding of how to improve their performance.

Frequency of Assessment

Researchers agree that formative assessments should not be done intermittently, but on a regular basis. In the literature on formative assessments, many authors point out the differences between various forms of tests that are given to students: high-stakes, interim, benchmark, exams, etc. (Black & Wiliam, 1998, 2009; Bloom et al., 1971; Hamilton, Halverson, Jackson, Mandinach, Supovitz, Wayman, National Center for Education, et al., 2009; Popham, 2008; Sadler, 1989). These are all types of assessment that are necessary, but different from what the authors look at when they talk about formative assessment. One of the biggest differences is frequency of the assessments. Formative assessments that measure student knowledge should be frequent, short and easy to administer, and used to complement summative data (Lembke, Stecker, & Center on Instruction, 2007). "Frequent assessment tasks, especially diagnostic tests, can help teachers generate cumulative information about students' levels of understanding and skill, so that they can adapt their teaching accordingly" (Nicol & Macfarlane-Dick, 2006, p. 211).

Despite these recommendations, it is unclear what is meant by the term "frequent." In a study by Fuchs et al. (2008) involving struggling 3rd grade math students, one of the key findings was that underperforming students need on-going progress monitoring, occurring about once per week. Much of the literature on special education and struggling students points to the need for more frequent progress monitoring of students. Ongoing progress monitoring is the most essential element in any type of intervention with low-performing students, and this needs to be in the form of formative assessments (Fuchs et al., 2008). Bangert-Drowns, Kulik, Kulik, and Morgan (1991) reviewed the effects of frequent class testing by conducting a meta-analysis on 40 studies. Their findings revealed that the more frequent the testing, the better students performed, but only up to a point; beyond about one to two tests per week, the performance would decline. In addition, they also found that several short tests had better outcomes than giving less frequent, but longer tests.

Formative Assessment Data Gathering and Use

In their assessment of multiple empirical studies of classrooms using formative assessments, Black and Wiliam (1998) found that student academic achievement can improve one-half to one full standard deviation. This implies that formative assessments are a must, but systems backed by research are not widely available, leaving schools to make the decision of whether to make their own or seek out and buy a pre-existing package (Sharkey & Murnane, 2006).

There are many different types of formative assessment, ranging from diagnostic tests to classroom questioning and checking for understanding during a lesson. Lembke et al. (2007) make the analogy of a formative assessment to a weekly health check-up that measures your blood pressure and temperature. No matter how often this information is gathered, it provides a snapshot into the students' current level of understanding. The difficulty lies in determining the instrument to use, what data to gather, and how the data will change the progression of instruction (Means, Padilla, DeBarger, Bakia, & Department of Education Office of Planning Evaluation Policy Development, 2009).

As was discussed above, only if the data gathered from the assessment are used to inform instructional decisions and provide feedback can the assessment be deemed formative. Finding systems that provide teachers with formative assessments to use with students that systematically provide data is difficult for both teachers and administrators (Goertz, Olah, Riggan, & Consortium for Policy Research in Education, 2009). Teachers, particularly in mathematics, have a difficult time determining how to collect and analyze the correct data in order to make instructional decisions, and therefore are in need of much professional development to refine these skills. Additionally, in order to change pedagogical practice, the formative assessment system must be sustained over long periods of time and continuously reinforced by teachers (Vendlinski, Hemberg, Mundy, Phelan, & Society for Research on Educational Effectiveness, 2009). In a randomized study on the impact of data-driven reform, Carlson, Borman, and Robinson (2011) found that even after data are gathered from various assessments, teachers struggle to access it from online databases and often have difficulty analyzing or using it to inform future decisions.

Multiple Choice Formative Assessments

Curriculum developers have to take into account what tools will help teachers facilitate learning without being over-burdening, but still provide ample information to make instructional decisions (Pellegrino, Winter 2002-03). This need gives rise to multiple choice (MC) formative assessments. In the argument for MC assessments, they can provide timely feedback and are easy to score, results are valid, and teachers are familiar with the methods and can analyze large amounts of data quickly (Sharkey & Murnane, 2006).

Creating MC formative assessments can be tricky and quite time consuming. MC questions imply one right answer and usually three or four wrong answer choices. These wrong answer choices, known as distractors, must be chosen carefully (Hedges, 1964). The creator of the question must have in-depth knowledge of common student errors and misconceptions, and use those strategically to formulate distractor choices so that the teacher is provided with meaningful formative data in order to make instructional decisions. If many students choose the same wrong distractor, then this provides information for the teacher about a possible misconception (Owens, Hanna, & Coppedge, 1970). For example, in an Algebra class, a question may be:

Distribute $-5(x-4)$

a) $-5x-20$ *b)* $-5x-4$ *c)* $-5x+20$ *d)* $-9x$

The correct answer here is *c.* It is very common for students to forget to distribute to all terms in the parentheses as in distractor choice *b,* or to not distribute the negative sign as in distractor choice *a,* or to combine all of the terms as in distractor choice *d.* In deciding what to do next, the teacher can address the mistake of the students, offering specific feedback on errors, and thereby correcting the problem. This kind of formative assessment gives both teachers and students information about learning and how to progress toward the learning goal (Sharkey & Murnane, 2006). The difficulty lies in the time, attention, and expertise that goes into creating these problem sets with specific distractor choices.

These types of MC formative assessments can also be powerful for students. They are able to receive clear (usually timely) feedback as to whether or not they scored correctly and which answer is correct. This gives them knowledge of the learning goal, feedback as to their performance, and the teacher can then give them instructions on how to get there (Tanner, 2003). Students are able to use metacognitive skills to analyze the reason why they got the question wrong – whether it was lack of knowledge, mistakes, or confusion. Students who answered correctly receive affirmation for their skills.

Checking for Understanding

Checking for understanding can take many forms from questioning, to summaries, to "exit slips," which are short open-ended questions that students use to show their learning from the day's lesson. Teachers also use reflection journals or Learning Logs as lesson-closing activities to have students reflect on what they have learned and possibly raise further questions (Fisher & Frey, 2007). These types of formative assessments are easy to administer, but the data are difficult to collect and analyze. Secondary teachers who see anywhere from 100 to 200 students per day would have a hard time collecting and sorting through detailed information like this on a daily or weekly basis. Therefore, choosing the appropriate type of formative assessment to use for the amount of students and data that are available is crucial for teachers to be able to provide targeted instruction, intervention, and make decisions about instruction.

The use of questioning during a lesson can provide students with instant feedback as to whether or not they are on the right track and can steer them in the right direction. Formative assessments in the form of written reflections, exit tickets, or summaries typically are best to supply the teacher with information rather than feedback for the students. Unless the teacher can provide written feedback for the students on an individual level, it is difficult for students to learn whether or not they are on the right track using this type of formative assessment. Sadler (1989) argues that when teachers provide a qualitative review of student work using multiple measures and provide evidence of expertise, then students will have the maximum opportunity to improve.

In a study of types of formative assessments used by teachers, conducted by Goertz et al. (2009), it was found that "if a teacher used *any* type of formative assessment…they were more likely to use *multiple* types of formative assessment" (p. 161, emphasis in original). Therefore, teachers do not typically use one form without another, but this statement also implies that there are teachers who do not use any type of formative assessment at all.

Data Analysis and Professional Development

Teachers can no longer wait until the end of a unit to give a test and only use summative data to evaluate students (Conderman & Hedin, 2012). Indeed, training on how students learn, how to develop assessments that measure learning, and how to interpret results should be mandatory for pre-service teachers, according to Pellegrino (Winter 2002-03). Pellegrino also argues that program accreditation for teacher licensure should include training in assessment.

It proves a challenge for many teachers to implement a formative assessment program because there is a lack of quality assessments available, and teachers often do not have the capacities to create their own (Phelan et al., 2011). Many do not have the deep content knowledge to build strategic questions, while others simply do not have the time. When deciding on formative assessment systems, teachers are better at incorporating assessments that are created for them versus having to create ones themselves (Frohbieter, Greenwald, Stecher, Schwartz, & National Center for Research on Evaluation Standards and Student Testing, 2011). This creates a tremendous need for formative assessment systems that provide programs of intervention which are backed by research (Black & Wiliam, 1998). While there is plenty of research that show the positive effects of formative assessments and feedback for student achievement, there are very few systems that are proven by research to be effective in multiple settings.

Researchers agree that formative assessment systems increase efficiency of learning in the classroom, but which programs to use and how to incorporate them into the already existing curriculum proves challenging for most teachers (Phelan et al., 2011). In order for formative assessment systems to work, there needs to be discussion amongst teachers and administrators as to which systems to adopt (whether they are already created or will be developed in-house), planning for implementation, and adequate training for teachers (Frohbieter et al., 2011). Follow-up trainings, professional development time, and on-going administrative support are all necessities. In addition, time should be set aside for teachers to discuss formative assessment results, analyze data, and share best practices. Teachers should look at assessment items one at a time, comparing their data to other teachers, and asking questions to make instructional decisions based on what works (Goertz et al., 2009).

Feedback

The topic of formative assessment cannot be discussed without incorporating the tightly coupled literature on formative feedback. Assessing students without providing feedback is not useful, and we should never test for the sake of testing. Indeed, if there is assessment without feedback, then it should be deemed summative (Black & Wiliam, 1998). "Formative assessment does make a difference, and it is the quality, not just the quantity, of feedback that merits our closest attention" (Sadler, 1998, p. 5). One of the key strategies of formative assessment according to Black and Wiliam (2009), is to provide feedback to students so that they may progress in their learning.

All too often, students take tests or quizzes with no goal (other than perhaps not to fail), and rarely receive timely feedback as to what they got wrong, why it was wrong, and what steps they can take to fix their errors. Therefore, in order for formative assessments to be effective, they must be tied to "clear criteria and high-quality feedback" (Phelan et al., 2011, p. 330). Feedback does not refer to praise or ridicule. The student needs to understand exactly where he/she went wrong and how to improve in order for feedback to be most effective (Hattie &

Timperley, 2007). The purpose of providing feedback to students is to help them increase their knowledge, skills, and understanding of the assessed item(s) (Shute, 2008).

Formative assessments should be used to help students progress in their learning, not just as a tool for teachers, and the best way to do this is through feedback (Nicol & Macfarlane-Dick, 2006). Exactly what constitutes feedback and what this looks like for teachers in the classroom is difficult for researchers to define. At a minimum, feedback has to serve a corrective function, indicate whether or not the student answered correctly, and provide information about the correct answer (Shute, 2008). Feedback has to strategically guide the student to be able to engage in meta-cognitive strategies to evaluate his/her performance against the learning objective. Telling a student “good job” or “try again” does not constitute effective feedback because the student does not have any specific information about how or why these comments were made (Clark, 2011).

Nicol and Macfarlane-Dick (2006) synthesized research on formative assessment and feedback to come up with seven principles of effective feedback practices. They concluded that good feedback practice:

1. helps clarify what good performance is (goals, criteria, expected standards);
2. facilitates the development of self-assessment (reflection) in learning;
3. delivers high quality information to students about their learning;
4. encourages teacher and peer dialogue around learning;
5. encourages positive motivational beliefs and self-esteem;
6. provides opportunities to close the gap between current and desired performance;
7. provides information to teachers that can be used to help shape teaching (p. 205).

Narciss and Huth (2004) outline another conceptual framework for developing formative feedback that is based on research of feedback types.

In their framework, Hattie and Timperley (2007) outline four levels of feedback that can be given at various times and situations to maximize effect:

- feedback about the task
- feedback about processing the task
- feedback about self-regulation
- feedback about self as a person.

They argue that feedback is most useful when it is specific to the task and about the process or how to progress. Giving a student feedback about himself is the least useful and can actually have the opposite desired effect. For example, telling a student he is not trying hard enough, or he can do better gives feedback about him as a person rather than specific information as to what he could do to improve his work. There are forms of positive and negative feedback, and each one works better with a different type of feedback. No matter what the type, according to Hattie and Timperley (2007), feedback must offer answers to three major questions:

- Where am I going?
- How am I going?
- Where to next?

The learning goals have to be specific (the where), the process or evaluation method must be clear (the how), and the student must have a clear idea of his progression (the next). When students feel that success is attainable, they are more likely to increase effort.

Timing of Feedback

Depending on the type of formative assessment, the timing of feedback can vary and be more or less useful. What remains ill-defined are the terms “immediate” and “delayed.” Each depends on the type of formative assessment being used. Immediate feedback could be as few as seconds, or as much as a day. Delayed feedback is typically thought of as at least a day or longer, giving time for the student to move on to other tasks/assignments before returning to the results of a previous one.

In a review of the literature of feedback via computer-based information systems, Mason and Bruning (2001) found that immediate feedback is most useful for low-performing students. This makes sense, since students who make many errors need frequent correcting and should not be allowed time to solidify incorrect procedures or misconceptions in their brains. In contrast, Mason and Bruning determined that high performing students benefitted more from delayed feedback, particularly on complex tasks. Hattie and Timperley (2007) outline the benefits of timing the feedback given in terms of the assessment type: for task-level assignments, they say delayed feedback is most beneficial, whereas for process-level assignments, it is best to have feedback given immediately.

The Art of Formative Assessment

The question of what exactly constitutes a formative assessment is not clearly answered in the literature. Essential components of formative assessment and formative feedback are broadly defined in relation to both the teacher and the student, but clear-cut examples of what forms this might take in various classrooms is sporadic at best. This could be because formative assessment remains, for the most part, a theory (Clark, 2010). It has definitions that are closely related, but there is no standardized practice of assessing in a formative way. Fundamental to the use of formative assessment is the flexibility of readjustment – once a teacher gathers data, the teacher has to determine next steps in order to enhance learning. No two data collections will reveal the same information, and two different teachers who assess their classes may not take the same course of action to intervene in student learning, setting each on a different path toward student achievement (Wiliam & Leahy, 2007). Therefore, the formative assessment must be looked at as a process, not necessarily a specific tool, and be seen as more of an art form than a science.

In his article on formative assessment, Clark (2010) finds it troublesome that many schools and districts seek packaged programs that offer set ways of collecting and analyzing data. Instead of using these programs, he argues that, as referenced in (Wiliam & Leahy, 2007), it would be much "more meaningful to practitioners and administrators to discuss the process in terms of the various functions that it actually serves" (p. 343). There is no specific method that is the best way to assess students that will work for every teacher across every content in every grade level (Black & Wiliam, 1998).

Doing What Works

The concept of using formative assessment data to improve teacher practice and student learning is very promising, but exceedingly difficult to carry out. Sharkey and Murnane (2006) outlined a few conditions that are necessary for successful formative assessment implementation based on an accumulation of research:

- assessments that are clearly aligned to state standards and to the curriculum being taught
- user-friendly technology that allows teachers with modest levels of computer literacy to explore patterns in test results

- teachers willing to learn how to analyze test results and how to derive implications for changes in instruction, and professional development to acquire the requisite skills
- time for teachers to do this work, ideally during the school day
- a school culture that embraces the idea that the achievement of all students is the responsibility of the whole teaching staff and that success depends on continued learning (p. 576).

Based on their study of a large school district, teachers had better success with formative assessment programs when they were able to create some of it themselves because it created "buy-in." Instead of being told what to do and how to do it, they had input as to what they wanted to do (Sharkey & Murnane, 2006). The bottom line is that formative assessment design is an art form, and many variables must be considered in order to find the right system that will work to improve teaching and learning.

Academic Vocabulary and Math Language

Teaching students academic vocabulary, which are the essential words or phrases for a content area, not only builds their background knowledge, but can increase their achievement by as much as one-third (Marzano, 2004). When students encounter word problems or have to engage in mathematical discourse, they often struggle with the vocabulary (Kovarik, 2010). Being able to speak math is a skill that many math teachers take for granted and neglect to teach explicitly to students. So although students can perform computations on paper, they cannot fully understand the concepts without knowing the language (Thompson & Rubenstein, 2000).

Rubenstein and Thompson (2001) emphasize the importance of building mathematics vocabulary and the language of math, which is the theory behind the use of translations in the DARTS model. Students have to practice translating mathematical phrases from English into math symbols in order to understand the vocabulary, learn how math is spoken/read, and understand concepts. Often overlooked when teaching academic vocabulary is the relevance of a vocabulary word to a concept (Marzano, 2004). Sometimes math vocabulary only makes sense when used in context. For example, the word *square* can be used as a noun in referring to the geometric shape or as a verb in referring to taking an expression to the second power. The same goes for the mathematical symbols. As stated by Rubenstein and Thompson (2001), “unless students develop this facility with symbols and the ability to recognize the desired use in a particular context, they will not be fluent users of mathematics” (p. 267).

Mathematics is like a language. Wakefield (2000) compares the following attributes of language with mathematics:

- Abstractions (verbal or written symbols representing ideas or images) are used to communicate.
- Symbols and rules are uniform and consistent.
- Expressions are linear and serial.
- Understanding increases with practice.
- Success requires memorization of symbols and rules.
- Translations and interpretations are required for novice learners.
- Meaning is influenced by symbol order.
- Communication requires encoding and decoding.
- Intuition, insightfulness, and "speaking without thinking" accompany fluency.
- Experiences from childhood supply the foundation for future development.
- The possibilities for expressions are infinite.

Correct vocabulary must be continuously practiced and enforced, because "if a student does not know how to read mathematics out loud, it is difficult to register the mathematics" (Usiskin, 1996, p. 236). One set of written symbols could be interpreted and pronounced many different ways. For example, $y - 12$ could be stated as "y take away twelve," "y minus twelve," "twelve less than y," "y less twelve," etc. It is only with deep knowledge of the meaning that students can connect the various word phrases with the symbolic representation (Rubenstein & Thompson, 2001). It is particularly important to pay attention to detail when speaking and writing math. In instances when order matters, students are prone to speak one thing and mean another. Seeing $5 \div 10$, students may say "ten divided by 5" or "5 divided into 10," and since a teacher knows what the student means to say, she lets it slide and does not feel the need to correct. Challenges that students face in the language of mathematics can be divided into three types: verbalization, reading, and writing (Rubenstein & Thompson, 2001).

Types of Academic Vocabulary

There are various types of academic vocabulary in mathematics. Some vocabulary words are very specific to the subject and are not otherwise used outside of the math class like *quotient* or *isosceles*. Hollingsworth and Ybarra (2008) refer to these types of words as content vocabulary, which are usually isolated to a specific content area. They also define academic vocabulary as those words which are interdisciplinary and are used throughout all content areas, such as *compare, contrast*, or *describe*. Then there are words that take on a different meaning when used in a mathematical context, like *mean* or *right* (Pierce & Fontaine, 2009). If a teacher orally asks the class "What is the sine of *x*?" without the context, students could misinterpret that as asking for the "sign of *x*." Subtleties of the language are often a point of confusion, especially for English Learners (ELs). For instance, "two less than a number" and "two is less than a number" differ only by the small word "is," which is typically taught to mean the equal sign. In this case, that small difference separates the concepts of $n - 2$ and $2 < n$ respectively. Neither expression uses an equal sign. That two letter word "is" can seem very insignificant to those not accustomed to the language, and while students who speak English are trying to just learn the language of math, non-English speakers have to learn how math is written and spoken in English (Martiniello, 2008).

Symbolic Representation of Words

When students read math, they cannot sound out the symbols if they do not know what the symbols mean. Not being able to mentally verbalize the symbols while reading math puts students at a great disadvantage and forces them to think only in the abstract. The implicit symbols that are "invisible" are often taken for granted and difficult for students to learn. The sign of a number is presumed positive if there is no sign in front of it. The coefficient of a variable is presumed to be 1 if there is no other number directly in front of it. When a term like $3x$ is written, the presumed operation is multiplication. Teaching students to know the hidden symbols is another difficulty in teaching the language of mathematics (Rubenstein & Thompson, 2001).

Student difficulties in writing math symbols happen for many reasons. Ask students to write "half of a number," and there are many responses that will come up. Not only are there various ways to write this expression using math symbols, but there are many more that are mathematically correct and yet not the most preferred way to symbolically represent the expression. For instance, a student may write $x \times \frac{1}{2}$, which is technically correct, but not the way a mathematician would write it. Students may also not understand the importance of very precise representation. It can be very challenging for students to evaluate an expression like

$\sqrt{b^2 - 4ac}$ using a basic calculator because they have to take into account grouping and the order of operations. So depending on the tool (pencil and paper or one-line calculator), students may have to record symbols differently (Rubenstein & Thompson, 2001).

Students do not often have chances to speak academically and discuss in the typical U.S. math classroom. Questions posed by the teacher are about the next step, the answer, or the computation. The use of mathematical vocabulary is typically done by the teacher during explanations, lecture, and questioning. Students follow along during the lesson using symbols on their paper and by listening to the teacher. They copy the notes, follow the steps, and find the solutions to the problems. Rarely are they asked to explain their reasoning in words or out loud. Students need the opportunity to speak, converse, and discuss their thinking, ideas, and questions in order to build their vocabulary as well as background knowledge (Marzano, 2004).

Over the past few years, 46 states have officially adopted the Common Core Standards – a set of standards for English and Math, grades K-12, that will help unite public education in America (NGA Center for Best Practices, 2010). The transition to the Common Core State Standards (CCSS) will require a fundamental shift in the way that mathematics is taught in the United States. Students will be expected to explain their reasoning, possibly using no symbols at all. There is a significant difference between a student who puts his steps into words and one who explains what he does mathematically. Students often “show their work” without being able to explain the mathematics. Sometimes they cannot even verbalize the symbolic representation correctly. Math teachers take for granted that if a student can show the steps by showing his work, then he must understand the concept, and subsequently, the mathematics behind it. Math education is focused on answer-getting. However, the answer should be only one part of the process, and learning mathematics is not about getting the answer using any means necessary (Daro, 2013). Math teachers try to get students to memorize rules: adding fractions, multiplying integers, dividing fractions, working with exponents. The problem is that students forget these rules and never understand the concepts involved. In fact, there are many teachers who cannot explain the mathematics behind the rules they were also taught in school. When asked why they invert and multiply when dividing fractions, some replied “That’s just the rule” (Molina, 2012).

Teaching students how to use shortcuts or gimmicks to find answers in mathematics can also hinder their understanding and ability to explain themselves. Some teachers are tempted to disguise math using fun sayings, mnemonic devices, and shortcuts. For example, referring to distribution as “the rainbow method” or the infamous FOIL (first, outside, inside, last – in reference to the order in which to multiply the terms and combine them together) method to explain multiplication of binomials. In his presentations to teachers and educational leaders, Dr. Phil Daro, one of the writers of the Common Core Math Standards, discussed what he has seen as “the butterfly method” for adding fractions with unlike denominators (Figure 1). The issue with these methods is that they have no conceptual foundation for students and do not emphasize the correct vocabulary.

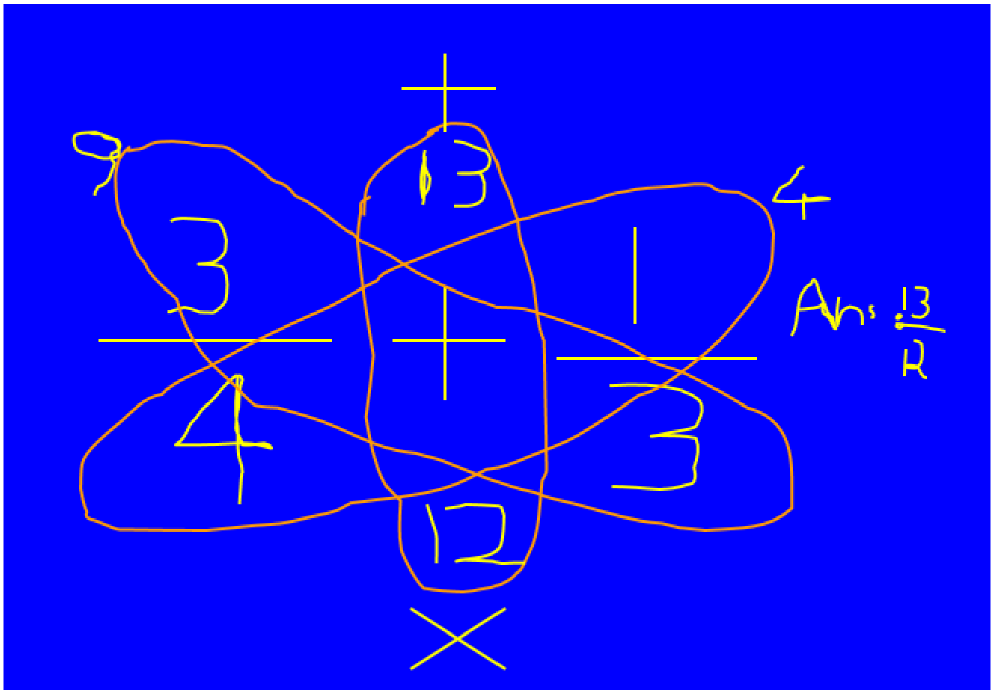

Figure 1 Butterfly Method. Reprinted from *CCSS Mathematics – A System of Courses*, by P. Daro, 2013, Retrieved from http://home.lausd.net/apps/news/show_news.jsp?REC_ID=331038&id=15.

Up until the adoption of the CCSS, the emphasis in math has been for students to learn facts, algorithms, and find the correct answer by any means necessary. Of course this led teachers to seek the quickest, most efficient and painless means by which to get students to memorize the shortcuts and get to the right answer. Rarely did questions ask for explanations or require students to know concepts behind mathematical facts. In turn, teachers of mathematics did not necessarily need to have deeper understandings of math either. But what the new CCSS assessments will ask students to do is explain their reasoning, create possible solutions within a set of boundaries, and use math vocabulary in ways that they have never been accountable for before (Tamayo Jr, 2010). Students must understand what the question is asking, what skills and concepts they will need to solve the problem, and be able to perform the mathematical concepts.

The need for vocabulary development is critical to develop math conceptual understanding, reasoning, and application to ensure success on these new assessments.

Word Problems

In a training for principals, Daro (2013) stated that the number one change that we must see in mathematics with the inception of the Common Core Standards is a shift away from answer-getting. He said that the American teacher looks at the math she needs to teach and says, "How can I get my students to get the answers to these problems?" while the teachers in other, more mathematically inclined (according to the TIMMS studies), countries say, "How can I use these problems to teach the mathematics." This shift is not going to be easy for American teachers whose textbooks are filled with answer-getting lessons and practice problems. It has long been known that both elementary and secondary math teachers will have to have a deeper knowledge of the mathematics (Ball, Lubienski, & Mewborn, 2001). They now also have to ensure that the learning does not stop with the correct answer to a problem; the learning should start with the answer.

If learning mathematics starts with the answers and uses problem-solving to increase student understanding of the concepts, then there is no better tool than word problems. Word problems are all about giving context to mathematics, and students need to apply their knowledge using context as often as possible (Greenwood, 2004). If math is how we make sense of many things in our world, why do we so often take our world out of mathematics? Molina (2012) describes this phenomenon as "naked numbers." Numbers are stripped of their meaning and context, and therefore become abstract ideas. When students struggle to understand a problem such as $7 - 3 = ?$, giving them manipulatives or a real-life example (i.e. You have 7 apples and give away 3. How many apples do you have left?) will help them get to the right answer. It is common for students to have trouble with arithmetic, and these students often use their fingers to help them count. This too helps give numbers context. So if putting context back into math helps students understand, the question remains: Why are word problems so difficult for students?

Student Difficulty with Word Problems

Difficulty with the language of math leads to difficulty with word problems. Not only do students need to understand the academic vocabulary, but the vocabulary must also be interpreted in the context of a scenario (Adams, 2003). Word problems are difficult both for students to learn and for teachers to teach. All too often, word problems are left out of math instruction, when really developing the problem-solving ability of students should be the most important goal of middle school mathematics (National Council of Teachers of Mathematics, 2000).

Having the appropriate background knowledge is essential to first understand the essence of the story within the problem (Hirsch, 2006). Only once this is understood can the student choose and apply the necessary mathematics to find a solution (Monroe, Panchyshyn, & Bahr, 2006). In many instances, if students can understand the word problem situation, they will seek to reason out an answer or use methods like guess and check in lieu of using an algebraic or algorithmic method (R. Hall, Kibler, Wenger, & Truxaw, 1989). For example, consider the problem: *Farmer John looks out onto his farm and counts 16 heads. The only animals are cows and chickens. He then counts 42 legs in all. How many of each animal does Farmer John have?* Young students (4^{th} – 7^{th} grade) can make guesses, draw pictures, or set up tables to figure out the quantity of each animal. Mathematically speaking, this is a system-of-equations algebra problem, which is not typically taught until at least the 8^{th} grade. Therefore, more complex math problems can sometimes be solved in easier ways through contextual associations, like word problems. Being able to contextualize helps students better see and understand how the mathematics work.

Nathan, Kintsch, and Young (1992) studied how students solved algebra word problems and the implications for teaching and learning. They concluded that students who are successful in solving these word problems are able to comprehend the problem informally as well as apply the formal mathematics. This often involves creating a connection from the concrete to the abstract. When teachers give students word problems in the context of a specific lesson in which the students have just learned a new skill, the word problem is relatively easy to solve. As long as the students apply the same skill they had learned that day, the problem-solving process is not overly difficult (Monroe et al., 2006). However, on a benchmark assessment, the word problems could involve any number of skills learned over the past few months, or even a combination of skills. Only when the students can make the connection between the story scenario and the skills that match can they be successful in solving the problem (R. Hall et al., 1989).

Another term for word problem is story problem. Fuchs et al. (2008) had this to say about story problems:

> *In contrast to number combinations and other aspects of calculation skill, story problems incorporate linguistic information that requires children to construct a problem model. Whereas a calculation problem is already set up for solution, a story problem requires students to use text to identify missing information, construct the number sentence, and derive the calculation problem for finding the missing information. This transparent difference would seem to alter the nature of the task and, in fact, some research suggests*

> *that computation and math problem solving may be distinct aspects of mathematical cognition* (p. 81).

The link between the problem-solving ability and the computational skills of a student may or may not be associated, as in the example of the Farmer John question above.

The CCSS call for increased rigor in mathematics by equally balancing procedural fluency, conceptual understanding, and application of the math standards. In their study of 5th and 6th grade students learning about decimal fractions, Rittle-Johnson, Siegler, and Alibali (2001) defined procedural fluency as "the ability to execute action sequences to solve problems" (p. 346). This type of mathematical knowledge can be applied without understanding the mathematical concept behind the problem or question. For example, a group of students is asked to represent $\frac{2}{3}$ (two-thirds) on the number line. Their choices were between 0 and 1, between 1 and 2, between 2 and 3, or more than 3. Most students divided to get a decimal representation, and answered between 0 and 1. Some students said it was between 2 and 3 (a common misunderstanding). One student explained that he divided 3 into 2 and got 6.67, which meant it was past 3 on the number line. This last student had some procedural knowledge in that he knew to divide, but misplaced the decimal point. Having no conceptual understanding of a fraction, he was unable to reason that his answer was incorrect. It is an iterative process that helps students develop a relationship between procedural knowledge and conceptual fluency (Rittle-Johnson et al., 2001), which is why the use of story problems in mathematics is so important.

Graphic Organizers

Graphic organizers come in various forms and are used for a variety of purposes. The use of a visual tool to assist with the bridge between understanding the word problem context and applying the mathematics is what Hyerle (1996) refers to as a task-specific organizer. T.

Hall and Strangman (2002) define graphic organizers as visual representations that depict the relationship between facts or ideas within a learning task. The Frayer Model, developed to present interrelated concepts, can be a useful graphic organizer for the purposes of solving a word problem (Frayer, Fredrick, & Klausmeier, 1969).

Having a graphic organizer for students to use when solving a word problem forces them to decipher relevant information, organize their thinking, develop a strategy, and choose an approach (Zollman, 2009). By using a graphic organizer, students can "organize and then clarify their thoughts, infer solutions to problems, and communicate their thinking strategies" (Zollman, 2009, p. 4). In a review of twelve studies that researched the effects of graphic organizers on learning by T. Hall and Strangman (2002), ten of those studies concluded a positive learning outcome with the use of graphic organizers. Within those studies, the most frequently studied learning outcome was comprehension. However, in a meta-analysis of 23 studies, which included both quantitative and qualitative studies by Moore and Readence (1984), findings suggest that the greatest learning outcome in the use of graphic organizers is vocabulary knowledge.

There are so many different kinds of graphic organizers, it is difficult to identify which are the most effective for a particular subject area or type of learning outcome. In a randomized experiment by Ives (2007), algebra students were divided into two random groups to learn how to solve systems of equations. One group received traditional explicit direct instruction and strategies, while the other received the same instruction through the use of graphic organizers. The group that used the graphic organizers outperformed the control group, and had a better conceptual understanding of systems of three linear equations as measured through a post-test. This particular study utilized students with learning disabilities; therefore graphic organizers help students who struggle with math to organize their thoughts and recall basic facts, rules, or procedures, which can extend to many other mathematical tasks (Strickland & Maccini, 2010).

Rubrics

Rubrics are generally used as tools for grading assignments or writing tasks that do not always have a specific correct answer, where process is just as important as product, a judgment of quality is required, or when the task is extensive and has many parts (Moskal, 2000). According to Panadero, Tapia, and Huertas (2012), rubrics have three parts: a list of necessary criteria, a scale by which these criteria are measured, and a description of quality at each scale level. Rubrics take much of the ambiguity out of grading and set clear expectations for students to follow. They are used at the conclusion of a task to measure quality, but it is recommended that rubrics be given to students before starting the task in order to help students meet all required criteria and eliminate guess-work as to what the teacher wants. Students should be able to self-assess by finding their own strengths and weaknesses using a criteria-referenced tool (Andrade & Valtcheva, 2009).

Rubrics help provide important feedback to students and give justification to teachers for giving a certain score. This can help students identify areas for improvement, but the rubric must be aligned with the purpose of the learning activity. A generic rubric to help a student improve a particular skill (i.e. problem solving or oral presenting) can provide students with specific feedback that can help them improve on subsequent tasks that are similar in nature (Moskal, 2000). In order for students to be able to meet the criteria in a well-developed rubric, teachers must create descriptions that are measurable and quantifiable for students (Brookhart, 1999). For example, “Essay contains no grammatical or spelling errors” would be easier to judge than “Paper is well-written,” which is more subjective to the grader.

The better a rubric is written, the greater the chance for inter-rater reliability. In other words, if a rubric's criteria are well-defined, and more measurable than subjective, the greater the chance that two separate teachers will come up with the same score for the task (Moskal, 2000). This can be particularly important in large schools where students may change classes and be assigned different teachers throughout the year. In many school districts, grades are left up to individual teachers, and can be highly arbitrary. Teachers decide what tasks they want to grade and how they want to grade them. Therefore, what constitutes an "A" grade in one class may be completely different for another class even within the same school. Having school-wide tasks that utilize the same rubric for scoring can help level-out the variations in grading.

Implications for Further Research

Implementing successful formative assessment programs in classrooms is a challenging charge for teachers. There is no lack of research about the effectiveness of quality formative assessment and feedback practices to increase student achievement. However, the research is typically grounded in a specific form of formative assessment and/or type of feedback practice. In general, research is lacking on systems of formative assessment and feedback loops; especially on a large scale. School or even district-wide formative assessment systems that are easy for teachers to use, provide sufficient feedback for students, and are flexible enough to work to meet all student needs are hard to come by.

There is insufficient research into what kinds of formative assessment work best for particular grade levels and content areas. One kind of formative assessment, such as multiple choice questions may work better in math, for example, while reflection and exit slips may work better for science. Different types of formative assessments that work well at the secondary level may not be appropriate for younger students. There is room for further research in specific areas of feedback as well. Additionally, identifying which forms of feedback work best when coupled with specific types of formative assessment is an area in need of further research. There are several feedback frameworks, so it would prove valuable for schools and teachers to know which one is backed by research to work best with the type of formative assessment plan they are using to help increase student achievement.

In the area of academic vocabulary, there are many articles showing the complexity of vocabulary in its various forms within mathematics, but a lack of research on the best methods for teaching students how to read, write, and speak the language of math. Vocabulary and word problems are closely coupled. How their instruction can be integrated is crucial to understanding how we can help students understand math concepts and apply their knowledge to the real world. With the adoption of the CCSS, there is room for ample research on the integration of the 8 Mathematical Practices, and what forms of instruction work best to increase student achievement on the new forms of assessment.

Finally, there is room for research on the integration of all of these research-based strategies for the use of intervention, closing the achievement gap, and increasing student achievement in mathematics. This study aims to take these four research-based, proven methods: formative assessment, feedback, academic vocabulary, and word problems, as a bundled program and test whether or not they are effective at increasing math test scores.

Summary

The research on formative assessments is in agreement that when used appropriately, there are positive outcomes for increasing teaching and learning in the classroom. Exactly how that looks in each classroom or school can be quite different. Some of the literature infers formative assessment data being mostly useful for teachers to make instructional decisions, while others feel it should mainly be used for students to develop metacognitive skills and adjust to meet learning goals. Many researchers believe that both students and teachers can benefit from good data.

Closely coupled with formative assessment is formative feedback. Feedback must provide students with clear information about the learning objective, what progress they are making, and how they can proceed to improve. There are many ways to collect formative assessment data and provide targeted feedback to students. The key is to find what works best for particular groups of students and measure effects through student learning gains. The research is clear – all classrooms should be using some kind of formative assessment process or tool to collect data in order for teachers to make instructional decisions and to provide students with valuable feedback in order to improve student achievement.

Academic vocabulary is essential for building a language of math. The research shows that there are different types of vocabulary used in mathematics, including content-based words, general academic words, context-dependent words, and mathematical phrases, all of which are important in learning this highly symbolic subject. Students are often taught short-cuts or algorithmic methods in order to learn how to get the answers to specific types of problems. However, the CCSS require students to be able to explain their reasoning, and they therefore must have not only a deeper understanding of the mathematical concepts, but a thorough knowledge of the math vocabulary in order to convey their thinking. Teachers must give students vocabulary instruction in addition to ample opportunities to read, write, and speak mathematically.

Word problems are the key to increasing the application of math. There is no lack of research on student difficulty with word problems, nor the benefits to student achievement when problems involving application are used frequently and consistently. These types of problems provide context for students. The use of graphic organizers can assist students in making sense of the problem, finding relevant information, and determining which math knowledge/skills to utilize. In order to decrease any ambiguity in the grading methods or expectations of solving a word problem, it is useful to use a rubric for scoring. Coupled together, a graphic organizer plus a rubric used for word problems can add consistency to the level of expectation and routine for students.

When combined into a curriculum, these components of formative assessment, feedback, academic vocabulary, and word problems, can have profound benefits for students. Previously researched in isolation, this study aims to discover their impact when used all together for the benefit of student achievement in mathematics.

Chapter III: Methodology

Research Design

This is a retrospective review of student achievement data, based on a secondary data analysis research design. The school which implemented DARTS with fidelity over several years (treatment school) was chosen as the focus for this study. Students from the treatment school will be matched with comparison groups from similar schools (schools with students similar in size, location, student scores and demographics) not implementing DARTS. This study will rely on multiple measures indicative of student achievement in mathematics, including: math achievement scores (scaled scores) on the CST, the Algebra 1 End-of-Course (EOC) Exam scores, and Academic Growth over Time (AGT) scores. The goal of this study is to demonstrate whether or not the DARTS math intervention model was effective in increasing student achievement in math. Internal data will be retrieved from the School District's Research Unit. Other data will be retrieved from public websites.

Sample

The sample will be a non-probability convenience sample consisting of the total population of students from the treatment school from 2007 through 2012, with the exception of students who were on an alternate curriculum or did not take the CST. The sample of students chosen for the match comparison groups will be the census of students who took the CST from two schools that were ranked as similar to the treatment school according to the 2009 data on demographics and Academic Performance Index (API) scores. In 2009, all three schools had 6th through 8th grade, and are in the same school district. Similar (non-treatment) School #1 was chosen based on its similar demographics and scores in 2009, but also because it had the same schedule (same number of math instruction minutes) as the treatment school from 2007 to 2009. Similar (non-treatment) School #2 was chosen for the same reason as #1, but also because it has the same schedule (90 minutes per day of math instruction) as the treatment school from 2010 to 2012. The target population of this study consists of 6th, 7th, and 8th grade students from the treatment school, located in a large urban school district ($n \approx 1{,}000$).

The demographics of students are outlined in Table 2 (data from http://api.cde.ca.gov). Although this population is not extremely diverse, it is representative of large schools in low-performing inner-city areas of California and other states close to the Mexican-American border. This population should be studied because of the complex needs these students face. They represent a growing population in the United States.

Table 2: 2009 Student Demographics

	Percent by School		
	Treatment School	Similar School #1	Similar School #2
Hispanic/Latino	97%	92%	92%
African American	3%	8%	8%
Socioeconomically Disadvantaged	99%	99%	98%
Students with Disabilities	11%	12%	11%
English Language Learners	40%	45%	40%

Protection of Human Subjects

No specific names of districts, schools, teachers, or students will be used in this study. Data will be gathered using pseudo-ID numbers for teachers and students. This study was approved after going through both the school district's IRB and Claremont Graduate University's IRB review because it uses existing data, retrieved and analyzed anonymously without information that could identify the students or teachers involved. This research is based on the longitudinal study of one school population and the comparison of students receiving or not receiving (from other schools) a particular curriculum. The DARTS treatment was developed by the researcher as part of a curriculum choice made by the school, unrelated to this study, beginning in the 2009-2010 school year.

Instrumentation

Scores from the California Standards Test (CST) for the 7th and 8th grades in all content areas will be gathered from 2007 through 2012 to look for gains and make comparisons. The CST is a criterion-reference test designed to assess student knowledge of the California content standards in English Language Arts (ELA), Mathematics, Science, and History – Social Studies. The math CST for all grade levels consists of 65 multiple-choice questions. In grades 6 and 7, these questions are aligned to a standard from one of the five content strands, as outlined in Table 3. For all CSTs, the cut score for the Basic level is always 300 and it is 350 for the Proficient level. Although these remain the same, the number of questions out of 65 that must be answered correctly to get that scaled score vary from year to year. Cut scores for the other band levels (Far Below Basic, Below Basic, and Advanced) vary by year, grade level, and content area. There is also a standard error of measure consisting of anywhere from 15 to 25 points around a particular cut score. The CSTs are checked for reliability and validity each year; the results of which are published in the California Standards Test Technical Reports (by year).

Table 3: Number (and Percentage) of CST Questions by Content Strand

	Number Sense	Algebra & Functions	Measurement & Geometry	Statistics, Data Analysis, & Probability	Mathematical Reasoning
6th Grade	25 (39%)	19 (29%)	10 (15%)	11 (17%)	Embedded
7th Grade	22 (34%)	25 (38%)	13 (20%)	5 (8%)	Embedded

Another instrument to be used is the Academic Growth over Time (AGT) reports by school. Beginning in 2010, schools in the district of this study received AGT reports by school, department, grade level, and subgroups. It reports one year at a time as well as a 3-year average. AGT is a value-added model that reports on a scale from 1 to 5, with 3.0 being average (district average). This report takes into account the prior academic achievement (based on CST scores) of students and makes a prediction as to how they should achieve based on their history and other demographic factors. It also compares like students with similar demographics to come up with a prediction of achievement. If the students achieve what was expected according to district averages, the AGT score would be 3.0. Any score under 3.0 is below the predicted growth, and any score exceeding 3.0 is above the predicted growth. As with all value-added models, there is a margin of error. Therefore, each score has a confidence interval, which shows the plus or minus margin of error given to that score. The score a school is given has a numerical score as well as a color score. The colors indicate the level of the confidence interval: red if the entire confidence interval is below 2.0, yellow if the entire confidence interval is below predicted growth, gray if the confidence interval crosses average growth, green if the entire confidence interval is above average growth, and blue if the entire confidence interval is above 4.0. The AGT score controls for the following variables: Prior CST Score, Grade Level, Gender, Race/Ethnicity, Low-Income Status, ELL (English Language Learner) Status, SPED (Special Education) Status, Continuous Enrollment, and Homelessness.

The Algebra End-of-Course (EOC) exam consisted of 40 multiple-choice questions and four free-response questions, of which the students must choose and complete two. Each multiple choice question counts for 2% of the total score, while the free response items are graded by teachers based on a rubric, and worth 10% each. Students must get 60% or better on the Algebra EOC to pass the exam. All of the questions were based on the 25 CA State Standards for Algebra 1. A passing score, along with a “C” or better in their Algebra 1 course was required for students to receive 10 units of high school credit for Algebra 1 and move on to the next course. Scores from the Algebra EOC will be reported as percent correct (out of 100%).

Treatment: The DARTS Model

All students in the treatment school participated in all five components of the DARTS model beginning the first week of the 2009-2010 school year. Each teacher was trained for approximately one hour during the first two days of school about expectations of implementation and how to use the data and assignments. Program fidelity was monitored using an implementation rubric and teachers were given feedback regarding their level of implementation and how to better follow implementation procedure (outlined below). Monthly meetings were held for math teachers to analyze data and student work, share best practices, and planning. The diagnostic, rescue assignment, translations, and story problems were done on a weekly basis with the exception of the first and last weeks of the semester and any week where students were taking a district math assessment or the CST. Assessments included district-wide assessments and the CST. The components of the DARTS model were implemented as follows in addition to the core curriculum at each grade level.

Diagnostic. The diagnostic is a 10-question multiple-choice (a-d) quiz given once every week in the math class. The same diagnostic is given for the entire grade/math level from SDC to Gifted. For lower grade levels, it typically takes no more than 20 minutes for students to finish the 10 questions. They may take longer on harder Algebra problems. The diagnostic should be given "cold" at the beginning of the period to gather what students genuinely know. Students should show all work, and the diagnostic should be given in a CST-like environment (monitored by teacher, no noise, no assistance, etc.). Questions should be as basic and skill-based as possible to begin. Distractor/wrong answer choices must be carefully chosen to target specific student errors, common misconceptions, and common student mistakes. The goal is to be able to infer what mistakes students are making when they chose a particular wrong answer.

When a large number of students choose one of the wrong answers, it is easier to point out and correct their mistake rather than reteach an entire concept.

With the exception of the first diagnostic given every year, each diagnostic is newly developed based on the previous week's school-wide data. Each question has a benchmark goal: 70%. If the school-wide data from a question reaches the benchmark, then that question can be considered "mastered" and is removed from the diagnostic. Another question in the progression of the learning of mathematics should replace it. However, if a question does not reach the goal, then that question will repeat on the following week's diagnostic (numbers should be changed). A question remains on the diagnostic until it reaches the benchmark. For example, a question might be $-9+5=$, and the answer choices are strategic to target common student errors in this type of problem. Even if the percent of students getting it correct begins at a very low percentage, that type of problem will stay on the diagnostic until it reaches 70%, and each week we are looking for growth. So the following week, that same question might become $-7+4=$.

Data Collection and Analysis. Teachers must collect and input the number of students who chose a, b, c, and d on each question and record it on the Google Spreadsheet. The spreadsheet takes those numbers and calculates the total responding, percent correct, average percent correct, best distractor, and develop graphs. Teachers should show the data for the students to see and analyze. There are data from the entire grade level, teacher level, and individual class level. Data are reviewed to see which questions the class did the worst on (for whole-class review) and specific distractors that many students chose to purposefully correct common mistakes.

Assessment. The DARTS model emphasizes the need for benchmark core-content assessments. The purpose of the diagnostic is intervention; it is NOT to test standards that were just taught. The DARTS model is in addition to (a layer on top of) the core math content. Therefore, it is beneficial to also have bi-monthly or quarterly benchmark assessments that do check the core content standards for the grade level. These assessments should be taken seriously, implemented rigorously, and data analyzed vigorously. Data from these assessments should be disaggregated by subgroups, monitored by administration, compared to previous years, and used to inform review and instruction moving forward.

Rescue Assignment. The rescue assignment is designed to match each week's diagnostic. It is differentiated intervention for students based on what each student needs to practice, providing instant feedback for what the student needs to master. For each of the ten questions on the diagnostic, there are 3-4 free-response questions on the rescue. Each student does the corresponding problems for each problem that s/he missed on the diagnostic. Some teachers will also assign problems on the rescue for which the student did not show enough work for on the diagnostic. Students are NOT to do the entire rescue assignment as "extra practice." That is not the design of the model. For example, if a student missed #2, 3, and 8 on that week's diagnostic, s/he would do #2, 3, and 8 on the rescue assignment, which would be about 12 extra homework questions. It is also a motivational tool – if a student gets a 10 out of 10 on the diagnostic, then s/he has no extra work to do, so it pays off to put in effort on the diagnostic.

Translations. Translations are math phrases completely written out in English words, which students translate into math symbols. In the younger grades an example would be *five plus six*, to which we would want the student to write $5 + 6$. Other examples could be: *the sum of a number and nine*, *twice the difference of two numbers*, *the perimeter of a square*, or *twenty nickels and dimes totals one dollar and sixty cents*. Students do not solve translations.

Teachers introduce 10 new translations per week, however they decide to divide them up (two in a warm-up each day, split in half over two days, or all at once). The translations are pre-developed and in sets that follow the progression of the grade level and build from literal translations to more abstract and conceptual understanding.

Translations are not just exercises – they are opportunities to engage in rich mathematical discussions utilizing proper vocabulary. A discussion about the ways to translate *half a number* engages students and can lead to a much deeper understanding. Possible answers should be shared and discussed, and teachable-moments should be utilized. The class should come to a consensus about which symbolic representation is “best” or “the most correct.” Correct mathematical vocabulary must be used by both the teacher and the students at all times.

Story Problems. A story problem is a word problem. The DARTS model requires students to engage with at least one story problem per week. This requirement can be met by doing one individually or in groups. Every story problem must follow the common (school-wide) graphic organizer and rubric.

The story problem is meant to build conceptual understanding of the core content. Use easy problems to get students used to using the graphic organizer and rubric, then go on to more challenging problems. Problems chosen should fit into the core curriculum, but can be differentiated to meet the various needs of students. Every classroom implementing DARTS must have evidence of story problems posted in the room. They should reflect the use of the graphic organizer and be graded by the rubric.

More information on the DARTS model can be found at www.karuza.com/darts.

Procedures

Data will be collected directly from the School District's Research Unit and entered into Excel and other statistical software, such as SPSS, for filtering and statistical analysis. Student demographic information and CST scores (by teacher) will be collected from the treatment school students in all content areas and all grade levels (6th, 7th, and 8th) each year from 2007 to 2012. Demographic information includes: gender, grade level, socioeconomic status, gifted/GATE status, English Learner Level, Special Education status, and ethnicity. CST scores include: scaled scores for all content areas tested (English (ELA), Math, History/Social Science (8th grade only), and Science (8th grade only)) and band level score (Advanced, Proficient, Basic, Below Basic, Far Below Basic) for all content areas tested. CST math scores and demographic information will be collected from the two similar non-treatment schools for matched-group comparison purposes. A list of teacher and administrator ID numbers each year will be requested in order to analyze changes in faculty from year to year. Finally, Algebra EOC exam scores for 8th graders will be requested from all schools (by teacher) between 2010 and 2012 (the only years the test was formally given).

Other public data will be gathered to further compare the impact of DARTS on the math program at the treatment school in comparison to the other two similar schools. This data will include: AGT Scores and other school-wide data available through the California Department of Education.

Proposed Data Analyses

1) *How did the DARTS math intervention model affect math achievement scores for students in an urban middle school from 2010 to 2012 in grades 6 to 8?*

In order to answer this research question, CST data and AGT data will be gathered for comparisons from year to year. Starting with data from 2007, the treatment school's CST data will be analyzed for 6^{th}, 7^{th}, and 8^{th} grade Algebra by looking at changes in mean scaled scores using t-tests, scaled score distributions, and percentage of students in each band level. Data from 2007, 2008, and 2009 will serve as pre-treatment (before DARTS was implemented) in order to determine any increasing, decreasing, or static trends. The 2010 data will show any changes after one year of implementation. The 2011 and 2012 data are after the school changed schedules and went to a double-block, giving all students 90 minutes of math instruction (in a core class or intervention class) each day. These two years will be analyzed in reference to each other.

2) *To what extent can outside factors (e.g. teachers, school leadership, school-wide initiatives, district reform) be excluded as reasons for changes in the treatment school's student achievement in mathematics during those years?*

Since this is not a randomly assigned experimental study, this research question mainly supports the validity of this study's findings. In order to help determine whether or not the DARTS model positively impacted student achievement, other outside factors must be accounted for as much as possible. One possible impact is an influx of new teachers or a new administration/principal. For instance, if a new administration comes in and makes systemic change throughout a school, there could be a significant difference in scores. The same could be said for teacher turn-over – if many new highly-effective teachers either joined or left the faculty, there could be a difference in scores. The following variables will be created and used in a correlation matrix to determine whether or not they had any relationship to student achievement scores: percentage of new teachers, whether or not there was a new principal, whether or not there was a new administration.

By gathering the CST data from other content areas (English, Science, and History-Social Studies) at the same school during the same years, changes in mean scaled scores, score distributions, and percentage proficient can be analyzed in comparison to math. School-wide initiatives would theoretically impact all content areas similarly, so this data can help answer whether or not changes in math achievement scores were directly due to the DARTS model.

The AGT data from each school year (beginning in 2010 when AGT first became available) is a value-added measure on a scale from 1 to 5. As stated in the instrumentation section, if the students achieve what was expected according to district averages, the AGT score would be 3.0. Therefore, if district reform expected all students to achieve at a higher level, then the 3.0 would represent that increase. Since AGT compares each school to the district average, higher scores in the treatment school can eliminate district-level reforms/influences on student achievement.

3) *What is the difference between math achievement scores of students in the treatment school who received the DARTS math intervention model compared to those from similar schools who did not implement the model?*

CST math scores will be collected for both the treatment school and the other two similar schools. This data analysis will be primarily done using two-tailed t-tests. The treatment school will be compared to a non-treatment school with the same schedule (same amount of time in math instruction per day), and this will vary depending on the year. In addition, similar data analyses will be done on the two non-treatment schools as was mentioned above in the first two research questions to show the difference between mean CST scaled scores and Algebra EOC scores, score distributions for both CST and Algebra EOC, percentage of students reaching proficiency on the CST or passing the Algebra EOC, changes in math scores in comparison to other content areas, and differences in AGT scores.

4) *Does the DARTS model have a greater impact on students in the treatment school who score in the low range (Below Basic and Far Below Basic), middle range (Basic), and/or high range (Proficient and Advanced) of the math CST?*

In order to answer this research question, analysis will be done on three groups of students: low (scoring Far Below Basic or Below Basic), middle (scoring Basic), and high (scoring Proficient or Advanced) ranges. AGT scores for the treatment school from 2010 through 2012 will be analyzed for the three groups to see if there are any differences. In addition, pivot tables (which count the number of students meeting two criteria at once) will be set up to analyze the change in CST band level (Advanced, Proficient, Basic, Below Basic, Far Below Basic) for one cohort of students moving from 6th grade in 2010 through 8th grade in 2012. This cohort will be divided into the three groups and the percentage change in scaled score will be analyzed through ANOVA. Additionally, a one-sample t-test will help determine if there were significant changes in mean scaled score within the groups from one year to the next. These analyses will help determine if DARTS makes a larger impact on students in a particular group.

Chapter IV: Results

The results of this study are presented in the following section. Each independent t-test was two-tailed, run only on valid CST scores (no modified assessment (CMA) scores and no blank scores were included). Attendance measures were controlled by filtering out students who were enrolled for less than two-thirds of the school year (< 108 days on a year-round calendar or

< 120 days on a traditional calendar). In addition, scores of students who had less than 95% attendance were not included in these tests.

Research Question #1

How did the DARTS math intervention model affect math achievement scores for students in an urban middle school from 2010 to 2012 in grades 6 to 8?

Tables 4 through 6 show the percentage of students scoring in each band level for the math CST.

Table 4: Treatment School 6th Grade Math– CST Band Level Percentages

	2007 N = 408	2008 N = 401	2009 N = 416	2010 N = 415
Advanced	1.2	2.2	3.8	9.4
Proficient	11.8	16.7	21.2	29.2
Basic	32.6	33.7	32.5	29.4
Below Basic	38.2	36.2	31.0	22.9
Far Below Basic	16.2	11.2	11.5	9.2

Table 5: Treatment School 7th Grade Math– CST Band Level Percentages

	2007 N = 433	2008 N = 331	2009 N = 388	2010 N = 348	2011 N = 404	2012 N = 449
Advanced	0.0	1.2	2.8	7.5	22.8	25.8
Proficient	12.2	21.1	11.6	20.4	30.4	28.7
Basic	27.7	29.0	34.8	32.8	25.8	24.9
Below Basic	40.2	31.4	33.5	28.7	16.9	13.6
Far Below Basic	19.9	17.2	17.3	10.6	4.2	5.8

Table 6: Algebra Treatment School 8th Grade– CST Band Level Percentages

	2007 N = 201	2008 N = 293	2009 N = 266	2010 N = 210	2011 N = 277	2012 N = 305
Advanced	0.5	0.3	0.4	1.0	15.2	19.7
Proficient	9.0	7.5	4.9	14.8	29.6	25.2
Basic	18.9	21.8	21.4	23.8	27.8	25.2
Below Basic	51.7	50.5	49.6	45.7	18.8	24.3
Far Below Basic	19.9	19.8	23.7	14.8	8.7	5.6

Independent sample t-tests were run for 6th grade, 7th grade, and 8th grade Algebra CST scaled scores in the treatment school between the following years: 2007 to 2008, 2008 to 2009, 2009 to 2010, 2010 to 2011, and 2011 to 2012. The first three tests were to establish any trends or baseline information. Table 7 shows the change in mean scaled scores for which tests were statistically significant with a Sig. (2-tailed) p-value of less than 0.05.

Table 7: Treatment school change in CST Math mean scaled score for statistically significant tests

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012
6th Grade	+9.8	+6.9	+18.8	N/A	N/A
7th Grade	+12.0	Not sig	+18.9	+40.7	Not sig
Algebra 1	Not sig	Not sig	+12.1	+51.5	Not sig

Between 2007 and 2008, the difference in mean scaled scores of both 6th grade ($p = 0.003$) and 7th grade ($p < 0.001$) were statistically significant. From 2008 to 2009, only the change in 6th grade scores ($p = 0.046$) was statistically significant, which is the year that the DARTS model was piloted in 6th grade.

Table 8: T-Test results for treatment school CST Math scaled scores 2009-2010

	Means	Std. Deviation	Sig.	t	Sig (2-tailed)
6th Grade	2009: 314 2010: 333	2009: 51 2010: 61	.001	-4.837	.000
7th Grade	2009: 302 2010: 321	2009: 46 2010: 56	.002	-4.972	.000
8th Grade Algebra	2009: 284 2010: 296	2009: 39 2010: 47	.001	-3.012	.003

The year that DARTS began school-wide, changes in mean scaled scores between 2009 and 2010 were statistically significant for all three grade levels: 6th grade ($p < 0.001$), 7th grade ($p < 0.001$), 8th grade Algebra ($p = 0.003$).

In 2011 the treatment school switched to a traditional calendar, adding an extra math tutorial class for all students (90 minutes of math per day), and became a 7th and 8th grade only school. Due to the increase in instructional minutes for math, data between 2010 and 2011 cannot be used to answer the research question, but are reported as follows. Between 2010 and 2011, both available grade levels also had statistically significant changes in mean CST scaled score: 7th grade ($p < 0.001$) and 8th grade Algebra ($p < 0.001$). Although there were small increases, none of the grade levels saw statistically significant changes in mean scaled scores between 2011 and 2012.

Research Question #2

To what extent can outside factors (e.g. teachers, school leadership, school-wide initiatives, district reform) be excluded as reasons for changes in the treatment school's student achievement in mathematics during those years?

The district was not able to provide information on teachers or administrators from the schools in this study. To help rule out school-wide initiatives, data from the other content areas were analyzed in the same way as math to see if there were similar results indicating some kind of school-wide impact.

The following tables show the score distributions (by percent) for ELA, Science, and History over the years in the treatment school.

Table 9: Treatment School 6th Grade ELA– CST Band Level Percentages

	2007 N = 408	2008 N = 401	2009 N = 418	2010 N = 415
Advanced	2.5	2.2	4.3	3.9
Proficient	12.3	18.2	20.6	22.2
Basic	30.4	39.7	36.8	40.7
Below Basic	30.9	23.9	25.1	20.7
Far Below Basic	24.0	16.0	13.2	12.5

Table 10: Treatment School 7th Grade ELA– CST Band Level Percentages

	2007 N = 433	2008 N = 355	2009 N = 389	2010 N = 348	2011 N = 412	2012 N = 474
Advanced	2.1	6.5	6.9	5.2	8.5	11.2
Proficient	16.4	18.9	21.1	20.7	29.1	27.8
Basic	32.8	29.9	34.4	39.1	32.8	34.2
Below Basic	32.3	25.6	20.1	17.5	16.5	17.1
Far Below Basic	16.4	19.2	17.5	17.5	13.1	9.7

Table 11: Treatment School 8th Grade ELA– CST Band Level Percentages

	2007 N = 473	2008 N = 409	2009 N = 396	2010 N = 316	2011 N = 409	2012 N = 412
Advanced	2.7	2.2	5.8	9.5	10.7	13.6
Proficient	11.8	17.6	16.6	17.1	23.0	25.5
Basic	37.6	37.7	36.3	35.1	35.5	29.5
Below Basic	26.8	28.1	26.9	21.5	20.3	14.6
Far Below Basic	20.9	14.4	14.1	16.8	11.0	10.2

Table 12: Treatment School 8th Grade Science– CST Band Level Percentages

	2007 N = 473	2008 N = 409	2009 N = 397	2010 N = 316	2011 N = 410	2012 N = 412
Advanced	3.0	8.9	11.6	18.4	15.4	26.7
Proficient	18.6	19.8	24.7	25.3	21.5	19.7
Basic	28.8	27.6	25.7	23.1	23.9	19.2
Below Basic	31.5	22.5	20.7	15.2	19.3	14.3
Far Below Basic	18.2	21.8	17.4	18.0	20.0	20.1

Table 13: Treatment School 8th Grade History– CST Band Level Percentages

	2007 N = 473	2008 N = 409	2009 N = 397	2010 N = 315	2011 N = 410	2012 N = 411
Advanced	1.3	3.7	4.3	6.7	8.5	10.5
Proficient	6.1	11.5	13.6	14.6	18.0	24.3
Basic	25.4	30.3	33.2	31.1	36.1	29.2
Below Basic	31.3	28.9	21.9	18.7	17.1	14.4
Far Below Basic	36.4	25.7	27.0	28.9	20.2	21.7

The treatment school CST scaled scores for content areas other than math were analyzed using independent sample (two-tailed) t-tests between each year for 6th grade, 7th grade, and 8th grade ELA, 8th grade Science, and 8th grade History.

Table 14: Treatment school change in CST English Language Arts mean scaled score for statistically significant tests

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012
6th Grade	+12.7	Not sig	Not Sig	N/A	N/A
7th Grade	+8.3	Not sig	Not Sig	+14.3	+7.7
8th Grade	Not sig	Not sig	Not Sig	+8.9	+9.2

Both 6th grade ($p < 0.001$) and 7th grade ($p = 0.024$) ELA had statistically significant changes in mean scaled scores between 2007 and 2008. None of the grade levels had statistically significant changes in mean scaled scores for ELA between 2008-2009 or 2009-2010. The treatment school began implementing a very similar model to DARTS in ELA during the 2010-2011 school year, including giving weekly formative assessments. Between 2010 and 2011 (7th grade: $p < 0.001$, 8th grade: $p = 0.019$) and 2011-2012 (7th grade: $p = 0.038$, 8th grade: $p = 0.012$), both grade levels had statistically significant improvements in mean scaled scores for ELA.

Table 15: Treatment school change in CST Science & History mean scaled score for statistically significant tests

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012
8th Grade Science	Not sig	+11.3	Not Sig	Not Sig	+24.5
8th Grade History	+14.8	Not Sig	Not Sig	+12.5	+9.1

Changes in mean scaled scores for 8th grade Science and History did not show any significant trends. Science had statistically significant increases in mean scaled scores from 2008 to 2009 (p

= 0.024) and from 2011 to 2012 ($p < 0.001$). History had statistically significant changes in mean scaled scores between 2007 and 2008 ($p < 0.001$), 2010 and 2011 ($p = 0.002$), and 2011 and 2012 ($p = 0.019$).

Beginning in 2010, the district began releasing Academic Growth over Time (AGT) scores. Generally, the scale is from 1 to 5, with 3.0 being average growth (the expected amount of academic growth was achieved in a year). Further explanation of how these scores are calculated along with their meaning was outlined in chapter 3.

Table 16: AGT Scores for the Treatment School

	2009-2010	2010-2011	2011-2012
6th Grade ELA	2.0	N/A	N/A
7th Grade ELA	2.7	4.0	3.9
8th Grade ELA	2.4	3.9	1.6
ELA Overall	2.4	4.0	2.9
6th Grade Math	3.8	N/A	N/A
7th Grade Math	2.8	5.6	6.1
8th Grade Algebra	2.3	5.1	2.4
Math Overall	3.1	5.8	5.7
8th Grade Science	N/A	2.0	1.2
8th Grade History	N/A	2.2	2.3

Math outscored all other content areas in AGT scores all three years, which were the years that the DARTS model was implemented. The scale is from 1 to 5, but several math scores were so far above the predicted growth, the AGT score exceeded the “highest possible” score of 5.

Research Question #3

What is the difference between math achievement scores of students in the treatment school who received the DARTS math intervention model compared to those from similar schools who did not implement the model?

Similar school #1 had the same bell schedule as the treatment school from 2007 through 2009, and serves as the comparison/control group for those years. Similar school #2 had the same bell schedule (same number of instructional minutes in math) as the treatment school from 2010 through 2011.

Table 17: 6th Grade Math Similar Schools – CST Band Level Percentages

	2007 N = 566	2008 N = 450	2009 N = 402	2010 N = 387	2011 N = 394	2012 N = 354
	Similar School #1				Similar School #2	
Advanced	0.9	4.4	3.2	9.0	13.7	11.9
Proficient	16.3	18.0	15.7	19.1	24.4	24.6
Basic	33.8	25.3	26.4	27.9	30.2	32.5
Below Basic	36.7	35.8	36.1	30.5	22.8	23.7
Far Below Basic	12.4	16.4	18.7	13.4	8.9	7.3

Table 18: 7th Grade Math Similar Schools – CST Band Level Percentages

	2007 N = 483	2008 N = 526	2009 N = 417	2010 N = 382	2011 N = 367	2012 N = 317
	Similar School #1				Similar School #2	
Advanced	1.0	2.7	1.0	1.8	4.9	7.6
Proficient	10.1	16.5	13.7	14.7	23.4	16.7
Basic	27.7	25.7	26.4	26.2	34.9	34.4
Below Basic	41.0	37.5	33.8	37.4	25.6	28.7

Far Below Basic	20.1	17.7	25.2	19.9	11.2	12.6

Table 19: 8th Grade Algebra Similar Schools – CST Band Level Percentages

	2007 N = 358	2008 N = 365	2009 N = 392	2010 N = 294	2011 N = 260	2012 N = 197
	Similar School #1				Similar School #2	
Advanced	0.0	1.4	1.0	0.7	1.9	4.1
Proficient	4.5	12.9	9.4	8.2	23.1	33.0
Basic	16.5	20.8	21.2	21.1	30.4	31.5
Below Basic	49.7	43.8	45.7	43.2	35.8	24.4
Far Below Basic	29.3	21.1	22.7	26.9	8.8	7.1

Independent sample t-tests were run for 6th grade, 7th grade, and 8th grade Algebra CST scaled scores in the similar schools between the following years: 2007 to 2008, 2008 to 2009, 2009 to 2010, 2010 to 2011, and 2011 to 2012. Table 20 shows the change in mean scaled scores for the similar schools (non-treatment schools) for which tests were statistically significant with a Sig. (2-tailed) p-value of less than 0.05.

Table 20: Similar Schools change in CST Math mean scaled score for statistically significant tests

	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012
	Similar School #1			Similar School #2	
6th Grade	Not sig	Not sig	+17.1	N/A	N/A
7th Grade	+9.3	-7.8	Not sig	+11.5	-5.1
Algebra 1	+11.6	Not sig	Not sig	Not sig	+18.5

These data do not show any trends of improvement for either of the similar schools. There was no year when all grade levels improved in math. At both schools, there were times of statistically significant decreases in 7th grade math scores.

In order to compare the treatment school with the corresponding similar school, an independent two-tailed t-test was run for each grade level each year. In 2007, the t-test revealed a statistically significant difference in the mean scaled scores for both 6th grade ($t = -2.196$, $p = 0.028$) and 8th grade ($t = 2.665$, $p = 0.008$). Similar school #1 had a higher mean scaled score ($M = 304$, $sd = 47$) than the treatment school ($M = 297$, $sd = 46$) in 6th grade. The similar school #1's mean scaled score for 7th grade also started out higher than the treatment school in 2007, but the differences were not statistically significant. For 8th grade Algebra, the treatment school had a higher mean scaled score ($M = 287$, $sd = 43$) than the similar school #1 ($M = 278$, $sd = 38$).

In 2008, the treatment school's mean scaled scores in math were nearly identical to similar school #1's in all three grade levels, but none of the differences were statistically significant.

In 2009, the year that DARTS began in 6th grade, the treatment school ($M = 314$, $sd = 51$) mean scaled score surpassed that of similar school #1 ($M = 402$, $sd = 54$) and the difference was statistically significant ($t = 3.379$, $p = 0.001$). Similar results were true for 7th grade scores: the treatment school ($M = 302$, $sd = 46$) was higher than the similar school #1 ($M = 294$, $sd = 48$), and the results were statistically significant ($t = 2.475$, $p = 0.014$). In 8th grade Algebra, similar school #1 had a higher mean scaled score than the treatment school, but the difference was not statistically significant.

When DARTS was implemented school-wide in 2010, the comparison of all three grade levels of the treatment school to similar school #1 yielded statistically significant results, with the treatment school's mean being higher.

Table 21: 2010 T-Test results for treatment school (T) and similar school #1 (SS#1) CST Math scaled scores

	Means	Std. Deviation	Sig.	t	Sig (2-tailed)
6^{th} Grade	T: 333 SS#1: 319	T: 61 SS#1: 63	.544	3.227	.001
7^{th} Grade	T: 321 SS#1: 298	T: 56 SS#1: 50	.056	5.839	.000
8^{th} Grade Algebra	T: 296 SS#1: 283	T: 47 SS#1: 46	.358	3.180	.002

Since the treatment school switched its bell schedule in 2010-2011, the 2011 and 2012 scores will be compared to similar school #2, which has always had that same schedule, giving students approximately 90 minutes of math per day. For background information, it is important to note that similar school #2 had higher mean scaled scores in math than the treatment school in all three grade levels in 2007, 2008, and all but 7^{th} grade in 2009 (although not all were statistically significant differences). These would be the years that similar school #2 had an additional 40+ minutes of math instruction per day as compared to the treatment school. In 2010, the differences in bell schedules between the two schools remained the same, yet still the treatment school out-performed similar school #2 with statistical significance in 7^{th} grade math ($t = 2.752$, $p = 0.006$).

The year that the treatment school switched schedules to match that of similar school #2 (2011), the comparison of all three grade levels yielded statistically significant results, with the treatment school's mean being higher. The same results occurred the next school year (2012). These results are outlined in Table 22 and Table 23.

Table 22: 2011 T-Test results for treatment school (T) and similar school #2 (SS#2) CST Math scaled scores

	Means	Std. Deviation	Sig.	t	Sig (2-tailed)
7th Grade	T: 362 SS#2: 322	T: 70 SS#2: 52	.000	9.085	.000
8th Grade Algebra	T: 347 SS#2: 314	T: 69 SS#2: 51	.000	6.379	.000

Table 23: 2012 T-Test results for treatment school (T) and similar school #2 (SS#2) CST Math scaled scores

	Means	Std. Deviation	Sig.	t	Sig (2-tailed)
7th Grade	T: 369 SS#2: 317	T: 82 SS#2: 56	.000	10.453	.000
8th Grade Algebra	T: 353 SS#2: 333	T: 78 SS#2: 54	.000	3.448	.001

Beginning in 2010, students enrolled in Algebra 1 took an End-of-Course (EOC) exam at the end of the year. Some of the similar schools had 7th graders enrolled in Algebra 1, but were not included in the statistical analysis since there were none at the treatment school. For 2010, the treatment school was compared with similar school #1, and in 2011 and 2012 compared with similar school #2, both using independent two-tailed t-tests. Scores were converted to be out of 100. In 2010, 26% of 8th graders from the treatment school passed the EOC with a 60% or higher, while 30% passed from similar school #1. In 2011, 67% of treatment school 8th graders passed the EOC, but only 31% passed from similar school #2. The treatment school had 51% of 8th graders pass the EOC in 2012 and 49% passed from similar school #2.

The only result that was statistically significant was in 2011 ($p < 0.001$) between the treatment school ($N = 277$, $M = 81.4$, $sd = 22.5$) and similar school #2 ($N = 339$, $M = 59.7$, $sd =$

21.4). In 2012, the treatment school had 18 more students than similar school #2, and had a higher average score, but it was not statistically significant.

In addition to these scores and tests are the AGT scores for all three schools in math for the years that DARTS was implemented at the treatment school.

Table 24: AGT Scores for the Treatment School, Similar School #1, and Similar School #2 in Math

	2009-2010			2010-2011			2011-2012		
	T	SS#1	SS#2	T	SS#1	SS#2	T	SS#1	SS#2
6th Grade Math	3.8	2.9	3.0	N/A	3.1	3.1	N/A	3.1	3.5
7th Grade Math	2.8	2.8	2.4	5.6	3.3	3.2	6.1	3.5	**
8th Grade Algebra	2.3	2.3	2.6	5.1	3.2	2.3	2.4	2.7	**
Math Overall	3.1	3.0	2.7	5.8	3.3	3.4	5.7	3.4	2.5

** Insufficient Data – No score available.

Table 25: AGT Scores for the Treatment School, Similar School #1, and Similar School #2 in ELA

	2009-2010			2010-2011			2011-2012		
	T	SS#1	SS#2	T	SS#1	SS#2	T	SS#1	SS#2
6th Grade ELA	2.0	2.9	3.1	N/A	2.2	3.1	N/A	2.4	4.0
7th Grade ELA	2.7	2.8	2.4	4.0	3.0	4.5	3.9	2.3	2.8
8th Grade ELA	2.4	2.3	3.2	3.9	2.1	3.2	1.6	1.8	2.6
ELA Overall	2.4	2.7	2.9	4.0	2.3	3.8	2.9	2.0	3.3

The overall math AGT score for the treatment school was higher than either of the similar schools all three years, which were the years DARTS was implemented. However, the overall ELA AGT scores were lower for the treatment school compared to both similar schools in 2010, above both similar schools in 2011, and in between the similar school scores in 2012.

Table 26: AGT Scores for the Treatment School, Similar School #1, and Similar School #2 in Science and History

	2010-2011			2011-2012		
	T	SS#1	SS#2	T	SS#1	SS#2
8th Grade Science	2.0	3.2	2.5	1.2	3.7	3.1
8th Grade History	2.2	2.7	2.0	2.3	1.8	2.6

**No data was available for 8th grade Science or History for the 2009-2010 school year.

For both years available, the treatment school's AGT score was below both of the other similar schools in 8th grade Science. The treatment school scored slightly above similar school #2 in 8th grade History in 2011, but slightly below in 2012.

Seven out of seven t-tests showed the treatment school significantly out-performed the respective similar school from 2010 to 2012. AGT scores were higher for the treatment school than the similar schools in 8 out of 10 comparisons for math, but only in 3 out of 14 for other content area comparisons.

Research Question #4

Does the DARTS model have a greater impact on students in the treatment school who score in the low range (Below Basic and Far Below Basic), middle range (Basic), and/or high range (Proficient and Advanced) of the math CST?

Table 27: Math AGT Scores for the Treatment School by Prior Student Achievement Level

	2009-2010	2010-2011	2011-2012
Advanced/Proficient	3.4	6.9	6.7
Basic	2.6	5.6	**
Below Basic/Far Below Basic	3.1	5.1	5.8

** Insufficient Data – No score available.

In the 2010 AGT report for the treatment school, the analysis scores are broken down by several types of demographic information, including prior level achievement for students. Students were divided into three groups: Advanced/Proficient (A/P), Basic (B), and Below Basic/Far Below Basic (BB/FBB). The A/P group scored 3.4 (above predicted), while the B group scored 2.6, and the BB/FBB group scored 3.1 (at predicted). In 2011, the A/P and B scores more than doubled to 6.9 and 5.6 respectively. The BB/FBB score went up to 5.1. Unfortunately, there was no score for the B group due to insufficient data in 2012. The A/P group scored 6.7 and the BB/FBB group had their highest score of 5.8.

By taking a cohort of students and following them from 2010 as 6th graders through 2012 as 8th graders enrolled in Algebra 1, data can be analyzed to see how these same students did over three years while participating in the DARTS model. Beginning in 2012, only those 8th graders with Algebra scores were pulled who also met the required attendance data. Out of those students, only those who had scores from 6th grade (2010 scores) and 7th grade (2011 scores), and met the attendance data requirement, were used, totaling 277 students. The students were divided up into three groups, depending on their level (low = BB/FBB, medium = B, high = A/P). Several tests were run on these groups.

Table 28: Math CST Placement Level Changes from 2010 to 2011 for Treatment School

	2011 Band Level (7th Grade)				
2010 Band Level (6th Grade)	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Far Below Basic		2	1		
Below Basic	1	9	23	9	
Basic		5	33	43	10
Proficient		1	6	49	48

Advanced				6	31

Table 29: Math CST Placement Level Changes from 2011 to 2012 for Treatment School

2011 Band Level (7th Grade)	2012 Band Level (8th Grade Algebra)				
	Far Below Basic	Below Basic	Basic	Proficient	Advanced
Far Below Basic		1			
Below Basic	4	10	3		
Basic	10	31	19	9	1
Proficient	2	26	38	35	9
Advanced		2	12	30	49

First, the math CST scaled scores of each group were analyzed with a one-sample t-test. The low group had a significant difference in mean scaled score ($p < 0.001$) from 282.5 in 2010 to 321.4 in 2011. The middle group also had a significant difference in mean scaled score ($p < 0.001$), increasing from 2010 (324) to 2011 (360.4). Another significant difference ($p < 0.01$) was found between the mean scaled scores of the high group as well, increasing from 396.5 in 2010 to 427.7 in 2011. The same students were arranged into the three groups again, but this time according to their 2011 score. One-sample t-tests were run for each group comparing their scaled scores from 2011 (7th grade scores) to 2012 (Algebra scores). There were only 18 students in the low group, but there was a statistically significant decrease ($p < 0.001$) in the means between years (284.6 in 2011 and 270.1 in 2012). The middle group also had a statistically significant decrease ($p < 0.001$) in mean scaled score from 329.3 in 2011 to 298.3 in 2012. The high group followed suit, showing a significant decrease ($p < 0.001$) from 416.8 in 2011 to 380.9 in 2012 in mean scaled score.

Taking this group's math CST scaled scores in 2010 and 2011, the difference was found and converted into a percent of change (difference in scores divided by the 2010 score). These percents of change were run through a one-way ANOVA and a Post-Hoc Scheffe test to see if there were significant differences between the groups. There was a significant difference ($F(2, 274) = 0.011$, $p = 0.02$) between the mean percent change in the high group ($M = 8.18\%$) and that in the low group ($M = 13.86\%$). The low group had a higher percentage increase. Running the same tests between 2011 and 2012 resulted in no significant differences between any of the groups.

Table 28 is a pivot table that shows the band level changes from 2010 (6^{th} grade score) to 2011 (7^{th} grade score) for these students. There were 49 students who started at Proficient and stayed there, while there were 48 students who started at Proficient and increased to Advanced the following year. Only 7 students dropped below the Proficient band level between 2010 and 2011. Table 29 is a pivot table that shows the band level changes of the same group of students from 2011 (7^{th} grade score) to 2012 (8^{th} grade Algebra score).

In each of the three years available, the high group had the most growth according to the AGT scores. The ANOVA which analyzed the percentage change in scaled scores indicated that the low group had the greatest percent increase from 2010 to 2011, and was significantly higher than the increase in the high group. Pivot tables show exactly how many students scored in each band level between years.

Chapter V: Conclusion

The purpose of this study was to demonstrate the effectiveness of the DARTS Math Intervention Model used by an urban middle school from 2009-2010 through 2011-2012. Results from the CST were used in t-tests to find statistically significant differences in mean scaled scores. The treatment school was analyzed against itself over the years as well as against two similar schools (similar demographics and bell schedules).

Research Findings

Research question #1: How did the DARTS math intervention model affect math achievement scores for students in an urban middle school from 2010 to 2012 in grades 6 to 8?

The DARTS model was first implemented in 6th grade only during the 2008-2009 school year, and went school-wide during the 2009-2010 school year. Sixth grade showed statistically significant increases in CST Math mean scaled scores beginning in 2008, and continued through 2010. Beginning in 2010, all three grade levels showed statistically significant increases in mean scaled scores (18.8 points for 6th grade, 18.9 points for 7th grade, and 12.1 points for 8th grade). In the percentage of students scoring proficient and advanced, 6th grade increased 13.6%, 7th grade increased 13.5%, and 8th grade Algebra increased 10.5%. This was the first year that all three grade levels showed significant increases in math scores.

The largest increases happened between 2010 and 2011. That was the year that the number of instructional minutes at the treatment school increased from about 53 to 90 per day. Therefore, the dramatic increases in scores, although statistically significant, cannot be solely linked to the DARTS model implementation. For further analysis of that particular year, the treatment school was compared with a similar school in research question #3.

Although both grade levels at the treatment school did show some improvement in mean scaled scores and percentage of students scoring proficient or advanced between 2011 and 2012, none of the changes were statistically significant. This means that the scores stayed relatively the same between the two years. What is also important to note during these years is the number of valid scores used in the mean. Between the years 2010 to 2012, the numbers in both 7th grade and 8^{th} grade increased significantly. The number of scores in 7^{th} grade went from 348 in 2010 to 404 in 2011 and to 449 in 2012. The number of scores used in 8^{th} grade Algebra went from 210 in 2010 to 277 in 2011 to 305 in 2012. This means that more student scores were included in the mean. For 8^{th} graders, not all students were placed in 8^{th} grade Algebra. The determination was based on previous scores, grades, and readiness for higher level math. Those who were deemed not ready entered an intervention course that repeated 6^{th} and 7^{th} grade standards. Typically, the more students in 8^{th} grade Algebra, the lower the average score. As the number of students rose in 8^{th} grade Algebra over the years, not only did the average not drop, but it continued to increase.

Research Question #2: To what extent can outside factors (e.g. teachers, school leadership, school-wide initiatives, district reform) be excluded as reasons for changes in the treatment school's student achievement in mathematics during those years?

The answers to this research question help validate the results of this study's findings, since this was not a randomized experiment. Initially, a correlation matrix was going to help determine whether or not large changes in faculty or administration had any relationship to the math achievement scores in the treatment school. However, the district was not able to provide that information.

The treatment school went through a number of school-wide initiatives, including the receipt of the Quality Education Investment Act (QEIA) grant, which gave the treatment school millions of dollars to reduce class size. Class sizes of core content (English, Math, Science, History) had to average at about 22.5 students, and none could exceed 27 or risk losing the entire grant. A school-wide change such as this could have impact on achievement scores, but in theory, would have similar outcomes in all content areas (English, Math, Science and History).

In other words, if the class size reduction helped improve student achievement in math, then it would also improve achievement in ELA, Science, and History. Therefore, numerous t-tests were run for each of the other three content areas between all available years. English Language Arts (ELA) had no statistically significant differences in CST mean scaled scores between 2008-2009 or 2009-2010. During those years, each student at the treatment school had two consecutive periods of ELA, totaling over 100 minutes per day. Recall from research question #1 that the changes between 2009 and 2010 for math (first year of DARTS implementation) were statistically significant for all grade levels. So although students had twice as much instructional time in ELA as they did in math, because of the DARTS model, students showed significant increases in math, but none in ELA during the same school year. Similarly, neither 8th grade Science nor 8th grade History had any significant increases that year. Therefore, it could not have been any type of school-wide initiative that caused the significant increases in math scores.

When the treatment school changed to a block schedule in 2010-2011, the number of minutes per day in ELA was reduced to only 90 (same as math). Despite this decrease in instructional minutes, the 7th grade mean scaled score increased 14.3 points and the 8th grade score increased 8.9 points, and both increases were statistically significant for the first time in three years. This was also the year that the ELA department began implementing a model similar to DARTS. The ELA model (termed DREAM) had a weekly diagnostic (formative assessment) and data collection, re-teaching mini lessons, editing (students fix grammar and punctuation), academic vocabulary, and model writing (a common writing prompt per grade level that the teachers would model). This means that extra time does not equate to a dramatic increase in scores, nor would a decrease in time equate to a decrease in scores. Additionally, the increase in ELA scores when they implemented DREAM, similar to that of math the year before, provide further support for the argument that formative assessments paired with feedback loops significantly impact student achievement scores.

Between 2011 and 2012, all content areas showed statistically significant increases in CST mean scaled scores except for math. Despite these increases, the mean scaled scores of the other content areas still did not reach the same level as either of the grade levels in math, nor did the percentage of proficient or advanced. In 2012, 7^{th} grade math had 54.6% of students scoring Proficient or Advanced on the CST and 8^{th} grade Algebra had 44.9%. Compare these percentages to 7^{th} grade ELA (39%), 8^{th} grade ELA (39.1%), 8^{th} grade Science (46.4%), and 8^{th} grade History (34.8%). Only 5.6% of 8^{th} graders in Algebra scored in the Far Below Basic range in 2012, compared to 20.1% in Science and 21.7% in History.

Starting in 2010, the district began releasing Academic Growth over Time (AGT) scores, which compare a student's actual academic growth (according to CST score) to the predicted growth. If the growth matched the predicted growth that should have occurred within that year, then the score would be 3.0 on a scale from 1 to 5. In the 2010 report, ELA AGT scores in the treatment school ranged from 2.0 to 2.7, and math AGT scores ranged from 2.3 to 3.8. The overall math score 3.1 was higher than ELA 2.4. The scores were expected to be higher in 2011 because of the schedule change, but math and ELA could now be compared because they had the same amount of instructional time per day. The overall ELA score (again, despite reducing the number of instructional minutes per day from over 100 to 90) was 4.0. The math scores were so far above the predicted growth that the score was above the range of the scale (from 1 to 5) at 5.8. Science and History both scored below the average at 2.0 and 2.2 respectively.

Although research question #1 showed that there were no significant increases in mean scaled score for either grade level in math in 2012, the AGT score of 5.7 shows that students still achieved far above predicted growth. Students achieved below the predicted growth average in ELA (2.9), History (2.3), and far below the predicted growth average in Science (1.2).

These scores mean that not only did students in the treatment school do much better in math during these years that DARTS was implemented, but they achieved above (and far above) the predicted growth as compared to similar students across the district. If there had been district-wide initiatives that improved student achievement, then all scores would have shown similar progress. Since students at the treatment school highly surpassed expected growth outcomes, it can be concluded that these changes in math achievement were not due to any district initiatives.

The findings from this research question further validate the results of the other research questions, demonstrating that the DARTS math model was effective in raising student achievement in math for students in an urban middle school.

Research Question #3: What is the difference between math achievement scores of students in the treatment school who received the DARTS math intervention model compared to those from similar schools who did not implement the model?

Independent two-tailed t-tests were run for the CST Math scores in the similar schools to compare mean scaled scores between years. Neither of the similar schools showed trends of increase for any of the grade levels. In addition, there was no year where all grade levels for one of the similar schools showed statistically significant growth. There was even one year per school that showed statistically significant decreases in student scores from one year to the next. Independent t-tests were run to compare the CST Math mean scaled scores of each grade level between the treatment school and the corresponding similar school. The mean scaled score for similar school #1 started out higher than the treatment school in 2007. They were about even in

2008, and the treatment school began to surpass the similar school in 2009. The most significant year was 2010, when the treatment school showed higher mean scaled scores in all three grade levels, all of which were statistically significant. Since this had not happened in any of the three years prior, it can be concluded that the treatment school out-scored the similar school because of the implementation of the DARTS model.

Looking at the Algebra EOC scores, there was only one instance where the treatment school had significantly higher scores than the similar school, which was 2011. The following year the averages on that test were much closer, although the treatment school still scored higher. Each year, more students took Algebra 1 and the EOC exam.

The treatment school was compared to similar school #2 for 2011 and 2012, since their bell schedules matched. Prior to 2009, similar school #2 had higher mean scaled scores in CST math than the treatment school (although not always significantly higher). In 2010, the treatment school began out-performing similar school #2 even with different bell schedules (treatment school had about 40 minutes less than similar school #2). T-tests revealed that the mean scaled scores from the treatment school were significantly higher than similar school #2 in both 7th grade math and 8th grade Algebra for both 2011 and 2012. Therefore all three years (2010, 2011, 2012) that the treatment school implemented DARTS, it had significantly higher CST Math mean scaled scores than the similar schools.

To further support these differences, this study looked at AGT scores for all three schools. In the 2010 report, all three schools were pretty similar in AGT scores; the treatment school had the highest overall math score of 3.1. The next year, in the 2011 report, all of the treatment school math scores were well above the "highest score" on the scale of 1 to 5, which were largely higher than even similar school #2. The treatment school had an overall math score of 5.8, compared to similar school #1's score of 3.3 and similar school #2's score of 3.4. The 2012 report gave the treatment school an overall math score of 5.7, while similar school #2 only scored 2.5. Scores from the AGT report for ELA, Science and History were also listed for all of the schools in this study in order to show that the treatment school was not out-performing similar schools in areas outside of mathematics, and in fact, often performed very similarly or lower. This information further demonstrates how the treatment school significantly out-performed both of the similar schools in math after implementing the DARTS model.

Research Question #4: Does the DARTS model have a greater impact on students in the treatment school who score in the low range (Below Basic and Far Below Basic), middle range (Basic), and/or high range (Proficient and Advanced) of the math CST?

Based on the scores from the AGT results, it appears that the DARTS model impacts all levels, but has a greater impact on students in the high range (Proficient and Advanced) of the math CST. Since the AGT scores take into account student prior year scores, it compares students to themselves, rather than what this study has done for the previous three research questions, which was to compare different groups of students. The scores of all three levels were fairly close, but in each of the three years from 2010 to 2012, the high group had the highest expected growth (AGT score) out of the three levels.

The scores of 277 students were used to track progress of a cohort of students from 6th through 8th grade at the treatment school who participated in the DARTS model for all three years from 2010 through 2012. These students were chosen because they met the attendance requirement all three years and took Algebra 1 in 2012. More students scored in the high group than in the low group as 6th and 7th graders. The one-way t-tests all had significant changes in mean scaled score, all showing increases from 2010 to 2011. The low group had the largest increase in mean scaled score the first year and the smallest decrease in mean scaled score the second year. It must be noted that the lowest group has the most room to increase while the high group has the most room to decrease. Additionally, the high group had the most students (N = 141 in the first test, N = 196 in the second test), so although the scores decreased the second year, there were more students in the sample and many students fell from Advanced to Proficient, but were still in the high group overall. In the ANOVA test where the three groups were analyzed for percent change between years, it appears that the low group improved significantly more than the high group from 2010 to 2011. However, this could be expected since the low group had greater room for improvement, while it is possible some students in the high group could not get much higher. The fact that the high group increased as much as it did is very impressive, since most intervention models only focus on the lowest group of students.

The pivot tables (Tables 28 and 29) show the change in band levels from one year to the next. Band level changes are difficult to analyze because it is ideal for students to move up in band levels, but it is also considered progress if students who start out in Proficient or Advanced remain in either of those bands (having a change of 0 or -1). Between 2010 and 2011 (from 6th to 7th grade), only 19 of the 277 students dropped a band level, 6 of which dropped from Advanced to Proficient. More students moved up one or two band levels that started out as Basic or Proficient than the others, indicating that the DARTS model was most beneficial to the middle and high groups.

From 2011 to 2012 there were a lot of students who dropped band levels. This could be because the standards and concepts in Algebra 1 are considerably more challenging and abstract than those in 6th and 7th grades. Still, 48% of the 277 students scored in the Proficient or Advanced band levels in 2012. According to Table 29, there was more movement in students moving downward in band levels than the previous years, but the highest numbers were those who stayed in the high level (remaining in Proficient or Advanced). Again, this indicates that the DARTS model most benefited those students in the high group.

It is unclear whether the DARTS model had a greater impact on the middle or lower levels in comparison to each other. According to the AGT scores, in 2010 the low group appeared to do better, but it was the high group who did the best the next year, and there was insufficient data to tell in 2012, but the high group did better than the low group. Looking at the changes in band level and percentages in each band level each year, it looks like students in the high group gained the most benefit from the DARTS model. However, it is important to note that the tests each year are not the same, and judging a student's score on the 7th grade CST compared to the Algebra CST is not perfectly calibrated. Clearly, all three groups showed marked improvement, and in 2011 and 2012 had above predicted growth. It is hard to say definitively whether the DARTS model was more beneficial to one level of students over another; this is an area in need of further research.

Limitations

The key limitation to this study is selection bias. Because there was only one school in this study that used the DARTS intervention model, there cannot be random assignment to the treatment, which means there is no formal control group. Therefore, only through matched comparison group analysis can this study determine causality. This study tried to account for as many confounding factors as possible. One of the threats to external validity is the sample's demographics. The students are nearly all Latino and socioeconomically disadvantaged. It may be challenging to make generalizable conclusions that apply to other ethnic groups of students.

Another limitation of this study is the test used to measure student achievement. Since this study largely relies on the results from the CST, findings may not be applicable to the new form of testing under the Common Core State Standards. Further studies will be necessary to research the impact that the DARTS model can make on student achievement under the new testing methods. The DARTS model itself may need to change to better prepare students for the cognitive demands of the new standards and tests. In order to demonstrate that the DARTS math intervention model is effective, it would have to be re-tested in a more current population sample with the Common Core Standards and assessments. Therefore, further studies need to be conducted.

The population sampling for this study was from former school years. Similar schools were chosen based on 2009 test scores, and these schools may not still be representable by that data sample. In addition, social/emotional morale could be different at various schools, affecting school cultures. These can have influence on student test scores and may be a threat to validity.

The ability to eliminate common rival explanations is difficult since schools and districts are always incorporating multiple initiatives to improve student achievement in parallel. However, the use of multiple measures and the ability to compare to other content area increases or schools within the same district will help alleviate this concern.

Finally, the DARTS model was developed and implemented at the treatment school by the researcher. It remains to be seen if results similar to the ones in this study can be replicated at another school site, run by someone other than the developer. It is possible that the researcher is able to implement the DARTS model better than anyone else. Further research could be done at other school sites to see if the DARTS model can be disseminated and replicated with the same level of fidelity and student results.

Implications

Overall, students at the treatment school did significantly better in math according to the CST than students in similar schools who did not implement the DARTS model. All three grade levels not only showed tremendous improvement, but sustained this consistent level of higher achievement. It is also important to note that, although this study is not about the achievement of students in ELA, when the treatment school implemented a similar model to DARTS in English Language-Arts, students did significantly better in that content area despite a reduction in instructional minutes. The implementation of frequent formative assessments helps teachers identify what students already know, how they are thinking what skills or standards they are struggling with, and what common mistakes or misconceptions need correcting (McIntosh, 1997). When done correctly, formative assessments can transform teaching and learning ("Formative assessment: Not just another test," 2011).

The DARTS model is more than a frequent formative assessment program. It is a comprehensive model that incorporates specific, targeted feedback in the form of targeted intervention, the academic language of the content, and problem solving. This implementation helped the treatment school begin to alleviate the nation-wide general trend of decreasing scores in the middle school grade levels in math (Alspaugh, 1998). The gap in achievement between 7th and 8th grade narrowed considerably in the treatment school. Vendlinski et al. (2009) state that formative assessment systems must be sustained over long periods of time and reinforced by teachers in order to be effective. Since the DARTS model is able to do this, in theory since each grade level achieves at a higher level, by the third year of going through the DARTS model students would perform highest in the 8th grade.

Data show that starting a model like DARTS in a school can significantly increase the number of students taking Algebra 1 and coming out successful after 8th grade. Early access to Algebra is a very significant indicator of how students will do in high school math (Green, 2012; Smith, 1996). Intervening early and providing support to struggling students can help them catch up to grade-level peers and make them more successful over time, rather than what has typically happened in secondary math, which is to decline as the grades increase. The more students are successful in middle school, the more they will be successful in high school (Wang & Goldschmidt, 2003), leading ultimately to a higher graduation rate and better success in college and career (Silver et al., 2008). The alarming statistic that students are only half as likely to graduate from high school if they have not passed Algebra 1 by the 9th grade (Silver et al., 2008) means that intervention in the middle grades for math is essential. Algebra has been known as the gateway to high school success, and DARTS helped the students at the treatment school achieve higher scores in Algebra than the students at similar schools.

Comparing the treatment school to the similar schools really showed the potential of the DARTS model. In both cases, the similar schools started out performing better in math than the treatment school. The treatment school began outperforming both similar schools after its first year of DARTS implementation in 2010. This has tremendous implications for math instruction because during the 2009-2010 school year, the treatment school did not have any “extra” time for math. There was no intervention period or block scheduling to give teachers additional time to incorporate the DARTS activities. Nearly all of the instructional time on Tuesdays was taken up by students taking the diagnostic and teachers collecting the data, leaving only four days to teach content and do DARTS. When instructional time is maximized and targeted, it can be put to greater use and results can exceed those from other schools who have more time, but do not use it efficiently. More time is not the cure, but quality time can make a large impact (Silva, 2012).

Similar school #2 has always had a block schedule, giving students nearly 90 minutes of math per day. When the treatment school changed to this schedule, student scores were significantly higher than those at the similar school in math. These findings are significant for math instruction because the DARTS model offers a curriculum with built-in intervention for schools that have math tutorial classes or intervention classes. Very often, students are placed in these classes arbitrarily, and the teacher has no information about the specific needs of the students. Weekly diagnostics given through the DARTS model provide teachers with necessary data to help students also pinpoint the “holes” in their math skills and knowledge to help build up their mathematical foundation. It gives teachers a systematic approach to gathering and analyzing data to target their instruction, while also working to build up students’ math vocabulary and problem solving skills. Through this type of intervention, students can then begin to access core content in their regular math classes, thus making tutorial and support classes effective in raising student achievement.

It is clear from the results of research question four that the DARTS model helped all levels of students improve achievement according to the math CST. Students in the high level (Proficient/Advanced) group showed the most above predicted growth according to the AGT scores and changes in band levels from 2010 through 2012. However, in 2011 and 2012, all three levels showed AGT scores that were far above predicted in math, which means that DARTS helped students at all achievement levels. There are many possible explanations as to why the high level group would benefit most from the DARTS model, and this would be an excellent area for further research. Students who already score Proficient or Advanced do not need a lot of intervention, but formative assessments in the form of diagnostics gave those students a lot of test-preparation practice, which helped them make less mistakes. It is possible that teachers cover more of the year's material with the higher level students, better preparing them for the CST test.

Nicol and Macfarlane-Dick (2006) found that formative assessment feedback loops can help students self-regulate their learning. Students who are already proficient in the standards have a clearer understanding of what they know and can more easily identify their mistakes, while students at a lower level may not know if it is their lack of understanding or a misconception/mistake that causes them to get problems wrong. Targeting specific student errors can help students in the low group find their specific gaps in learning, enabling them to further access the grade-level content and catch up at an accelerated pace. Students in the low levels automatically have more "ground to cover" in order to catch up to their grade-level peers. A comprehensive model, like DARTS, helps meet their specific intervention and skill-based needs while also teaching them the vocabulary and problem-solving skills of mathematics.

All of the results from research questions one through three help demonstrate the effectiveness of the DARTS model in raising student achievement in math as measured by the CST. When compared to its past scores prior to 2009, the treatment school showed tremendous growth in all grade levels during years when DARTS was implemented, and then sustained that growth in 2012. To add validity to the results, this study analyzed the other content areas in the treatment school to eliminate any school-wide strategies or initiatives like class-size reduction. Results in the other content areas did not come close to those found in math. Using both t-tests and the AGT scores, other content areas did show some growth over time, but not in the same way as math during the years that DARTS was implemented. To further show its effectiveness, it was important to compare the treatment school to similar schools that did not implement the model. In the years prior to 2009, both similar schools outscored the treatment school. After the implementation of DARTS, the treatment school showed significantly higher scores through both t-tests and AGT scores than either of the similar schools in math. All of these results indicate that the DARTS model is quite effective in increasing student achievement in math in an urban middle school.

Recommendations

Middle schools need comprehensive intervention models that can be implemented in conjunction with any curriculum. The models must fill the learning gaps for students, and most importantly, must be affordable and feasible for schools to use. Using frequent formative assessments to collect data and use the information to drive instruction and intervention is vital for improving student achievement. Whether using state standards or the new Common Core Standards, the need for formative assessment data does not change. Gaps in student learning and skill sets will always exist. The question is what will be done to alleviate these gaps?

Teachers need access to frequent formative data. This need also means that teachers must be provided with training and support on how to analyze data. Even when data are readily available, using data to make instructional decisions continues to be challenging. Data are not the reason that schools improve; rather, it is what is done in response to data outcomes that help reform schools. Unfortunately, school districts and teacher credential programs do not typically focus on developing the skills that teachers and administrators need in order to properly analyze student data (Means et al., 2011). Educators are left to "look at the data," – a collection of numbers, percentages, and graphs – without any follow-up as to what it means or what needs to be done about it (Hamilton, Halverson, Jackson, Mandinach, Supovitz, Wayman, National

Center for Education Evaluation, et al., 2009). There must be adequate training for teachers about how to read the data, what it means, and why it matters to them and their students. The DARTS model allows for teachers to record, analyze, share, and effectively use data on a weekly basis to drive their instruction.

It is necessary for schools to have systems in place to address a multitude of student needs. Some of these needs include poor math skills and a lack of mathematical understanding. English Learners need a way to not only acquire English, but the language of math as well (Martiniello, 2008; Molina, 2012). It is highly recommended that schools have a comprehensive model in place, such as the DARTS model, to address all of these needs simultaneously. It is not sufficient to place all struggling students into a class and remediate them with the constant re-teaching of fractions and decimals. Students need to engage in meaningful mathematics, helping them apply what they learn by speaking, writing, and problem solving (National Council of Teachers of Mathematics, 2000).

In order to get students ready for Algebra (whether that be in 8th grade or in high school), intervention needs to happen early for those students who begin falling behind. Math is a subject that inherently builds on previous skills (National Governors Association & Council of Chief State School Officers, 2012). As students matriculate to middle school, intervention needs in math must be identified early. Waiting until 8th grade to give students extra intervention time during the school day is too late. Many districts wait until the student is programmed into Algebra before providing a "support" class. It would behoove schools to start this process in 6th and 7th grade.

Theoretical Implications

The success of the DARTS model as described in this study is largely due to the formative assessment and feedback-loop of the diagnostic and rescue assignment. In particular, the feedback and what is done with the data that are collected each week help both the teachers and students know where to go next with instruction and learning. The theoretical implication of this study supports the findings in the article by Hattie and Timperley (2007) on the Power of Feedback. “Feedback allows students (and/or their teachers) to set further appropriately challenging goals as the previous ones are attained, thus establishing the conditions for ongoing learning” (Hattie & Timperley, 2007, pp. 8-9). This study shows how powerfully proper feedback can impact student learning and achievement, particularly in middle school math.

Feedback has to be the right kind, and the DARTS model helps teachers provide that kind of progress-oriented guide to next steps for students to figure out their gap in skills, mistakes, and misconceptions. Math teachers are able to use data no more than a week old to target student intervention needs for their individual classes, allowing for better differentiation (providing different services for students with different needs). The implications of the formative assessment/feedback-loop can be far-reaching, working in many different content areas. The treatment school took the DARTS idea and was able to successfully adapt it to English Language-Arts. There is no reason why the DARTS model could not be adapted to work in Science and History as well, since it is mainly based on the feedback theory of Hattie and Timperley. Further research could be done on similar models in different content areas.

Summary

The DARTS model incorporates formative assessment, feedback and individualized intervention, academic vocabulary/language, and problem solving to layer the complex work of providing intervention with meeting grade-level cognitive demands. Without any of these pieces, the model would not be as effective; each one helps support the others. The fact that the year the treatment school began implementing a similar model in English Language-Arts gave them very similar growth speaks to validity of the design of programs such as DARTS. There is something to be said for a model that can be implemented across all grade levels in a large urban school and show tremendous student achievement results.

Overall, this study showed that the DARTS model was successful in raising student achievement in mathematics at the treatment school. Various schools are now using the DARTS model in their math intervention classes. It is flexible and adaptable. However, in order to implement the DARTS model with fidelity, there is a lot of work and commitment that must go into creating the diagnostics and rescue assignments. Going forward, it would be ideal to work on a way to stream-line this process in order to alleviate the burden on personnel to create the parts, leaving them with more time to analyze data and work on planning lessons. It can have a tremendous impact on the math instructional program, providing a high level of consistency throughout the school and across grade levels. Although this study was in an urban middle school, it could feasibly work well in high school and non-urban schools (even those with very high performing students). There is no limit to where this model could potentially improve student achievement.

References

Adams, T. L. (2003). Reading mathematics: More than words can say. *The Reading Teacher, 56*(8), 786-795. doi: 10.2307/20205297

Alspaugh, J. W. (1998). Achievement loss associated with the transition to middle school and high school. *The Journal of Educational Research, 92*(1), 20-25. doi: 10.2307/27542186

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19.

Association for Supervision and Curriculum Development (ASCD). (1996). *Developing performance assessments. Facilitator's guide*. Alexandria, VA: ASCD.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. *Handbook of research on teaching, 4*, 433-456.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Barber, B. K., & Olsen, J. A. (2004). Assessing the transitions to middle and high school. *Journal of Adolescent Research, 19*(1), 3-30. doi: 10.1177/0743558403258113

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education, 5*(1), 7-74.

Black, P., & Wiliam, D. (2006). *Inside the black box: Raising standards through classroom assessment*: Granada Learning.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 1*(1).

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formation and summative evaluation of student learning*. New York: Mac Graw-Hill.

Brookhart, S. M. (1999). The art and Science of classroom assessment: The missing part of pedagogy. Eric digest.

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. *Formative classroom assessment: Theory into practice*, 43-62.

California Department of Education. (2013). 2012 star test results. Retrieved April 28, 2013

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*(3), 378-398. doi: 10.2307/41238557

Clark, I. (2010). Formative assessment: 'There is nothing so practical as a good theory'. [Article]. *Australian Journal of Education (ACER Press), 54*(3), 341-352.

Clark, I. (2011). Formative assessment: Policy, perspectives and practice. *Florida Journal of Educational Administration & Policy, 4*(2), 158-180.

Cohen, D., & Hill, H. (2000). Instructional policy and classroom performance: The mathematics reform in California. *The Teachers College Record, 102*(2), 294-343.

Conderman, G., & Hedin, L. (2012). Classroom assessments that inform instruction. *Kappa Delta Pi Record, 48*, 162-168. doi: 10.1080/00228958.2012.733964

Daro, P. (2013). Ccss mathematics: A system of courses. Retrieved from http://home.lausd.net/apps/news/show_news.jsp?REC_ID=331038&id=15

Fisher, D., & Frey, N. (2007). *Checking for understanding: Formative assessment techniques for your classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

Formative assessment: Not just another test. (2011). *Education Digest: Essential Readings Condensed for Quick Review, 76*(9), 16-19.

Frayer, D. A., Fredrick, W. C., & Klausmeier, H. J. (1969). *A schema for testing the level of concept mastery*: Wisconsin Univ. Research & Development Center for Cognitive Learning.

Frohbieter, G., Greenwald, E., Stecher, B., Schwartz, H., & National Center for Research on Evaluation Standards and Student Testing. (2011). Knowing and doing: What teachers learn from formative assessment and how they use the information. Cresst report 802: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Fuchs, L. S., Fuchs, D., Powell, S. R., Seethaler, P. M., Cirino, P. T., & Fletcher, J. M. (2008). Intensive intervention for students with mathematics disabilities: Seven principles of effective practice. *Learning Disability Quarterly, 31*(2), 79-92. doi: 10.2307/20528819

Goertz, M. E., Olah, L. N., Riggan, M., & Consortium for Policy Research in Education. (2009). From testing to teaching: The use of interim assessments in classroom instruction. Cpre research report # rr-65: Consortium for Policy Research in Education.

Green, K. V. S. (2012). *The critical middle years and the relationship of early access to algebra on high school math course completion and college readiness.* West Virginia University Libraries.

Greenwood, S. (2004). Content matters: Building vocabulary and conceptual understanding in the subject areas. *Middle School Journal, 35*(3), 27-34. doi: 10.2307/23044238

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6*(3), 223-283. doi: 10.1207/s1532690xci0603_2

Hall, T., & Strangman, N. (2002). Graphic organizers. *National Center on Assessing the General Curriculum. [Online] URL: http://www.cas.org/ncac. (accessed: 29/7/2012)*.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., National Center for Education, E., Regional, A., & What Works, C. (2009). Using student achievement data to support instructional decision making. Ies practice guide. Ncee 2009-4067: National Center for Education Evaluation and Regional Assistance.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., National Center for Education Evaluation, Regional Assistance, & What Works Clearinghouse. (2009). Using student achievement data to support instructional decision making. Ies practice guide. Ncee 2009-4067: National Center for Education Evaluation and Regional Assistance.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 1-32. doi: 10.3102/003465430298487

Hedges, W. D. (1964). How to construct a good multiple-choice test. *The Clearing House, 39*(1), 8-11. doi: 10.2307/30180309

Hirsch, E. D. (2006). *The knowledge deficit: Closing the shocking education gap for American children* (Kindle ed.). New York: Houghton Mifflin.

Hollingsworth, J., & Ybarra, S. (2008). *Explicit direct instruction (edi): The power of the well-crafted, well-taught lesson*: SAGE.

Hyerle, D. (1996). *Visual tools for constructing knowledge*: ERIC.

Ives, B. (2007). Graphic organizers applied to secondary algebra instruction for students with learning disorders. *Learning Disabilities Research & Practice*, *22*(2), 110-118.

Kovarik, M. (2010). Building mathematics vocabulary. *International Journal for Mathematics Teaching and Learning*.

Lembke, E. S., Stecker, P. M., & Center on Instruction. (2007). Curriculum-based measurement in mathematics: An evidence-based formative assessment procedure: Center on Instruction.

Los Angeles Unified School District. (2012). All youth achieving: Performance meter: Author.

Martiniello, M. (2008). Language and the performance of english-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in schools*: ASCD.

Mason, B. J., & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. *Retrieved February, 15*, 2007.

McIntosh, M. E. (1997). Formative assessment in mathematics. *The Clearing House, 71*(2), 92-96.

Means, B., Chen, E., DeBarger, A., Padilla, C., Department of Education Office of Planning Evaluation Policy Development, & SRI International. (2011). Teachers' ability to use data to inform instruction: Challenges and supports: Office of Planning, Evaluation and Policy Development, US Department of Education.

Means, B., Padilla, C., DeBarger, A., Bakia, M., & Department of Education Office of Planning Evaluation Policy Development. (2009). Implementing data-informed decision making in schools: Teacher access, supports and use: US Department of Education.

Molina, C. (2012). *The problem with math is english: A language-focused approach to helping all students develop a deeper understanding of mathematics*: Wiley. com.

Monroe, E. E., Panchyshyn, R., & Bahr, D. L. (2006). Helping children succeed with word problems. *Mathematics in School, 35*(1), 4-5.

Moore, D. W., & Readence, J. E. (1984). A quantitative and qualitative review of graphic organizer research. *The Journal of Educational Research, 78*(1), 11-17.

Moskal, B. M. (2000). Scoring rubrics: How? : ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *Timss 2003 international mathematics report: Findings from iea's trends in international mathematics and Science study at the fourth and eighth grades*: ERIC.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 181-195).

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction, 9*(4), 329-389. doi: 10.1207/s1532690xci0904_2

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics* (Vol. 1). Reston, VA: Author.

National Governors Association, & Council of Chief State School Officers. (2012). Common core state standards initiative Retrieved June 16, 2012, from www.corestandards.org

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards mathematics*. Washington, D.C.:

National Governors Association Center for Best Practices, Council of Chief State School Officers.

NGA Center for Best Practices. (2010). Common core state standards initiative Retrieved November 30, 2011, from www.corestandards.org

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Owens, R. E., Hanna, G. S., & Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement, 7*(2), 87-90. doi: 10.2307/1434339

Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), 806-813.

Pellegrino, J. (Winter 2002-03). Knowing what students know. *Issues in Science and Technology*, *XIX*(2), 48-52.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13. doi: 10.1111/j.1745-3992.2009.00149.x

Phelan, J., Choi, K., Vendlinski, T., Baker, E., & Herman, J. (2011). Differential improvement in student understanding of mathematical principles following formative assessment intervention. *The Journal of Educational Research, 104*(5), 330-339.

Pierce, M. E., & Fontaine, L. M. (2009). Designing vocabulary instruction in mathematics. *The Reading Teacher, 63*(3), 239-243.

Popham, W. J. (2008). *Transformative assessment*. VA: ASCD.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346.

Rubenstein, R. N., & Thompson, D. R. (2001). Learning mathematical symbolism: Challenges and instructional strategies. *Mathematics Teacher, 94*(4), 265-271.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*(1), 77-84.

Sharkey, N. S., & Murnane, R. J. (2006). Tough choices in designing a formative assessment system. *American Journal of Education, 112*(4), 572-588. doi: 10.1086/505060

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi: 10.2307/40071124

Silva, E. (2012). Off the clock: What more time can (and can't) do for school turnarounds. *Education Sector*.

Silver, D., Saunders, M., & Zarate, E. (2008). What factors predict high school graduation in the Los Angeles Unified School District *California Dropout Research Report #14* (Vol. 14): UCLA and UCSB.

Smith Green, K. V. (2012). *The critical middle years and the relationship of early access to algebra on high school math course completion and college readiness.* Doctor of Education, West Virginia University, Morgantown, WV.

Smith, J. B. (1996). Does an extra year make any difference? The impact of early access to algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis, 18*(2), 141-153.

Strickland, T. K., & Maccini, P. (2010). Strategies for teaching algebra to students with learning disabilities: Making research to practice connections. *Intervention in School and Clinic, 46*(1), 38-45.

Tamayo Jr, J. R. (2010). Assessment 2.0:"Next-generation" comprehensive assessment systems. *An analysis of proposals by the partnership for the assessment of readiness for college and careers and smarter balanced assessment consortium: Aspen Institute*.

Tanner, D. E. (2003). Multiple-choice items: Pariah, panacea, or neither of the above? *American Secondary Education, 31*(2), 27-36. doi: 10.2307/41064484

Thompson, D. R., & Rubenstein, R. N. (2000). Learning mathematics vocabulary: Potential pitfalls and instructional strategies. *Mathematics Teacher, 93*(7), 568-574.

Usiskin, Z. (1996). Mathematics as a language. *Communication in mathematics, K-12 and beyond, 86*.

Vendlinski, T. P., Hemberg, B., Mundy, C., Phelan, J., & Society for Research on Educational Effectiveness. (2009). Designing professional development around key principles and formative assessments to improve teachers' knowledge to teach mathematics: Society for Research on Educational Effectiveness.

Wakefield, D. V. (2000). Math as a second language. *The Educational Forum, 64*(3), 272-279. doi: 10.1080/00131720008984764

Wang, J., & Goldschmidt, P. (2003). Importance of middle school mathematics on high school students' mathematics achievement. *The Journal of Educational Research, 97*(1), 3-17.

What Works Clearinghouse. (2004). Curriculum-based interventions for increasing K-12 math achievement--middle school. What Works Clearinghouse topic report. Washington D. C.: What Works Clearinghouse.

What Works Clearinghouse. (2007). Middle school math. What Works Clearinghouse topic report: What Works Clearinghouse.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, *22*(5), 537-548. doi: 10.2307/1501668

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. *J. McMillan, H.(ed.) Formative Classroom Assessment: Theory into Practice*, 29-42.

Zollman, A. (2009). Students use graphic organizers to improve mathematical problem-solving communications. *Middle School Journal, 41*(2), 4-12. doi: 10.2307/23044804