DISSERTATION

USING A VARIATION OF THE COHORT CONTROL DESIGN TO EVALUATE LARGE-SCALE, LONG-TERM, COMPLEX PROFESSIONAL DEVELOPMENT PROGRAMS

Submitted by

Laura Sample McMeeking

School of Education

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2009

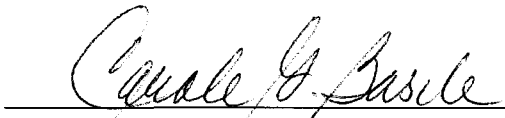
ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

COLORADO STATE UNIVERSITY

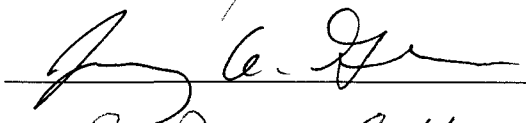
May 13, 2009

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY LAURA SAMPLE MCMEEKING ENTITLED USING A VARIATION OF THE COHORT CONTROL DESIGN TO EVALUATE LARGE-SCALE, LONG-TERM, COMPLEX PROFESSIONAL DEVELOPMENT PROGRAMS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate work

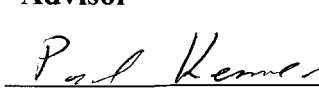


Carole Basile

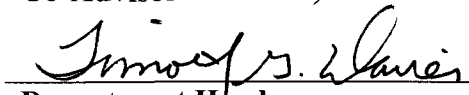


Jeffrey Gliner

Advisor R. Brian Cobb



Co-Advisor Paul Kennedy



Department Head Timothy Davies

ABSTRACT OF DISSERTATION

USING A VARIATION OF THE COHORT CONTROL DESIGN

TO EVALUATE LARGE-SCALE, LONG-TERM,

COMPLEX PROFESSIONAL DEVELOPMENT PROGRAMS

The purpose of this study was to introduce a variation on the posttest-only cohort control design and answer questions concerning both methodological credibility and practical utility of employing the variation design in evaluations of large-scale, complex professional development programs. The original design and variation design, which adds a pretest measure for prior student performance, were compared theoretically and practically using data from the RM-MSMSP program to compare the advantages and disadvantages of the two evaluation designs.

Two separate 2 x 2 ANOVA analyses were used to compare the designs. Findings indicated that, as expected, there were differences in the outcomes using the two designs. While the outcomes were not consistently different, they could plausibly be explained. Because the findings of the variation design were supported by similar findings in the literature, credence was given to the variation design. Given the added control of the variation design, evaluations using the variation design could control for selection bias where those using the original design could not. Realistically, however, the choice of evaluation design is one of trade-offs, because the addition of controls through gain scores, as is the case for the variation design, also comes with some disadvantages. If

certain conditions for the data are met and the outcome measure is calibrated, the variation design would be a good choice for a professional development evaluation. If not, the original design would also be perfectly reasonable.

From a practical standpoint, the variation design is no less practical to employ than the original cohort control design if certain program conditions about data collection and availability are met. In addition, the outcome measure in the form of a gain score is similar to value-added evaluation designs that are politically popular due to a focus on student growth. This coupled with the methodological advantages of the variation design make it a useful evaluation design for large-scale, long-term, complex professional development programs wishing to investigate the effects of professional development on student achievement.

Laura Sample McMeeking
School of Education
Colorado State University
Fort Collins, CO 80523
Summer 2009

ACKNOWLEDGEMENTS

Many people have been played a role in the successful completion of this project. These mentors, associates, and friends have believed in me, supported me, and given me numerous bits of advice throughout this whole process. I would especially like to express my gratitude to my advisor, R. Brian Cobb, for the never-ending support he has given me, the confidence he had in my abilities to complete this work, and his trust in my crazy ideas to get the job done.

I would also like to thank Paul Kennedy, who gave me the opportunity to be a program evaluator in the first place. Without his belief in my ability as a program evaluator, I would never have found this project. My thanks also go to Carol Basile who allowed me to move far away and still be a part of this program evaluation. I am especially appreciative to Jeffrey Gliner for supporting the direction of this dissertation. Without him, this dissertation would either be much longer or on a completely different topic.

I cannot forget to thank my husband, Gavin, who had a special insight to the pressure of finishing this dissertation and was successful in the impossible feat of keeping me sane throughout the dissertation process. I am forever blessed to have parents who have supported and encouraged me my entire life and in-laws who love me as their own daughter. Above all, I give thanks to God who has given me everything for nothing in return.

For my husband and family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 – INTRODUCTION

The mathematical achievement of students in the United States has been a major concern for the past two decades. In 1989, Carpenter, Fennema, Peterson, Chiang, and Loef published their randomized professional development study, which found that professional development could improve student achievement. Encouraged by similar findings (e.g., Kennedy, 1998), sweeping educational reforms came in 2001 in the form of the No Child Left Behind (NCLB) Act, which supports professional development aimed at improving teacher quality and boosting student achievement (NCLB, 2001). In addition, funding organizations that support professional development began focusing on projects whose goals were tied to student achievement. For example, the National Science Foundation (2006) created the Math Science Partnership initiative, which creates partnerships between universities and K-12 institutions with the goal of developing and implementing methods to increase student achievement in mathematics and science.

Still, by 2005, the National Center for Education Statistics (2006) reported that fewer than one-third of American eighth grade students performed at the "proficient" level, a level that defines a basic competency of challenging subject matter. They also found that as many as one-fifth of fourth graders and one-third of eighth graders lacked the skills to execute basic mathematics computations, which was more troubling. With all the investment in high quality professional development aimed at improving student achievement, why are students not achieving at higher levels?

Wayne, Yoon, Zhu, Cronen, and Garet (2008) assert that while professional development studies may show improvement in teacher quality and student achievement, they give no clear guidance in successful professional development characteristics that span across different contexts and settings. The lack of consensus most likely stems from the complexity of evaluation professional development programs. Due to the complexity of professional development programs, many evaluations are no more than observational studies, while others attempt quasi-experimental or even true experimental designs. However, just as professional development program leaders are faced with design challenges such as which program characteristics to include, professional development evaluators are faced with design challenges such as which evaluation design best fits the program while still maintaining rigor and decreasing bias. This is often difficult given the complexities of professional development programs, which are often non-random and long-term with teachers moving in and out of the program at different points in time. Investigating the link between teacher professional development and student achievement adds another level of difficulty, as the evaluator must somehow match student data with teacher data to get at this link between teacher professional development leading to teacher practice leading to student achievement.

Purpose of Study

The Rocky Mountain Middle School Math and Science Partnership (RM-MSMSP) was an NSF-funded professional development program that sought to address a gap in the literature concerning middle level (grades five through nine) student achievement related to teacher professional development in math and science (National Mathematics Advisory Panel, 2008; Hill, 2007; Mohr, 2006). The hypothesis was that

teachers who are exposed to challenging mathematical content would have increased basic content knowledge, be more confident in their teaching, and ultimately increase student mathematics achievement (M. Jacobson, personal communication, October 28, 2008). The purpose of this study was to compare two methodological approaches to evaluating the RM-MSMSP program with the intent of improving the causal claims of the evaluation, which was meant to determine the effects of the professional development program on student achievement.

Given the complexity of the RM-MSMSP program, the variation on the posttest-only cohort control design as defined by Shadish, et al. (2002) was created to enhance the credibility of the RM-MSMSP program evaluation's causal claims. The posttest-only design and variation design presented in the next chapter illustrate the problems associated with quasi-experimental research designs in relation to causal claims, especially given the complexities of professional development evaluation in general. While there are obvious issues, there are also ways to mitigate the bias associated with non-randomization. It is often impossible in these large programs to employ random assignment, so a strong quasi-experimental design, such as the two designs compared in this study, is necessary for evaluating project outcomes. With proper design elements, the problems related to quasi-experimental design can be mitigated, and more confident causal statements can be made.

Using theoretical and experiential knowledge, this comparison study aimed to determine the feasibility of employing this variation design in the evaluation of a large, non-randomized, complex professional development program. This research will add to the field of study in two ways. First, it provides an alternative quasi-experimental design

for use in other large, non-randomized studies. The variation design could be used by other researchers and program evaluators to aid in formative and summative evaluations. Second, it provides a discussion about the methodological and practical implications of employing such a design for large, non-random, complex professional development programs. While there are several obvious limitations to this study, which are discussed later in this chapter, the overall findings will inform other program administrators and evaluators about the advantages of employing either of the two compared designs and how to utilize the variation design for their own professional development programs.

Evaluator's Perspective

This study was of personal interest to me as both an evaluator and a scientist. The process of creating a new research design (or at least modifying an existing design) and testing its validity and utility was fulfilling as it allowed me to use my logical, process-oriented skills. From a practical standpoint, having seen how useful mathematics can be in everyday life, it is important to me to pass on good mathematical knowledge and skills to children so that the United States can compete in both the technical and non-technical job markets. Therefore, in conducting this study, I hoped to validate a new evaluation design that could later be used to investigate what elements of professional development we can use to help teachers better affect student learning.

As a program evaluator, I tend to view the world through a pragmatic lens. Creswell (2007) describes a pragmatist as one who is most often concerned with the practicality of research; attention is paid to applications of the research and solutions to problems. In subscribing to this paradigm, I believe there is more than one way to research something and choose to use the most practical methods available for any given

study. However, as I am a scientist by training, the pragmatism I inherently subscribe to is grounded in post-positivism. Therefore, the designs and methods I typically employ in research usually take on logical elements and employ empirical data collection that can lead to causal statements. This study is not an exception, as the evaluation design being introduced is quantitative in nature and employs a logical representation of the links between different pieces of the professional development and student achievement.

Research Questions

The research questions asked in this study come from the formation of the evaluation plan, as it became obvious that in order to infer causality, it might be necessary to create and use a new design. Therefore, the main question being answered in this study concerns the creation of the new design: does this design offer a new methodological solution to the problem of practice? In other words, does this design allow evaluators to adequately investigate the effects of teacher professional development on student achievement? To help answer the main question, four sub-questions were asked that address the credibility and usability of the variation design needed to be asked.

1. How do the outcomes using the variation design compare with those of the original cohort control design?
2. How credible are the causality statements inferred by the outcomes of the variation on the cohort control design?
3. What conditions must be met by the program in order for the design to have merit?
4. How practical is the variation design for use in programs similar to the RM-MSMSP program?

Theoretical Framework

The theoretical framework of this RM-MSMSP evaluation was based on the theory of change evaluation model. Chen (1990) defines a theory as "a set of interrelated propositions with the purposes of explaining and predicting a phenomenon" (p. 40). The theory of change model requires program implementers to specify what they believe will happen as a result of their intervention. In essence, the implementers test or create theory surrounding the links between program interventions and program outcomes. The NSF encourages program implementers to analyze the theory underlying their programs, and by doing so, better account for intervention effects. However, critics argue that understanding program theory is not a substitute to the empirical measurement of program outcomes and the comparison of these measurements with counterfactuals (The Brookings Institution, 1998). In fact, Rogers (2000) notes that few programs have attempted to address the measurement of causal linkages defined by the theory of change. This study addresses these criticisms by introducing and validating a new evaluation design that can be used to analyze the some of the theoretical links laid out by RM-MSMSP program implementers (see Figure 1).

Figure 1 expresses the theoretical linkages between the professional development, teacher learning, and student achievement that were tested in the full RM-MSMSP evaluation. It also includes outside effects such as teacher characteristics (e.g., teaching experience) and student characteristics (e.g., student ethnicity, student socioeconomic status, and student disability status). Under this framework, teachers who attend the RM-

MSMSP professional development have increased content knowledge and utilize reformed instruction in their classes. This teacher knowledge and pedagogical reform will affect student achievement. At the same time, teacher and student characteristics are also affecting student achievement. The double sided arrows in Figure 1 depict that the professional development, knowledge gains, pedagogical reform, and other effects are happening simultaneously. Overall, Figure 1 shows that professional development, teacher characteristics, and student characteristics all affect student mathematics achievement at the elementary and middle school level.

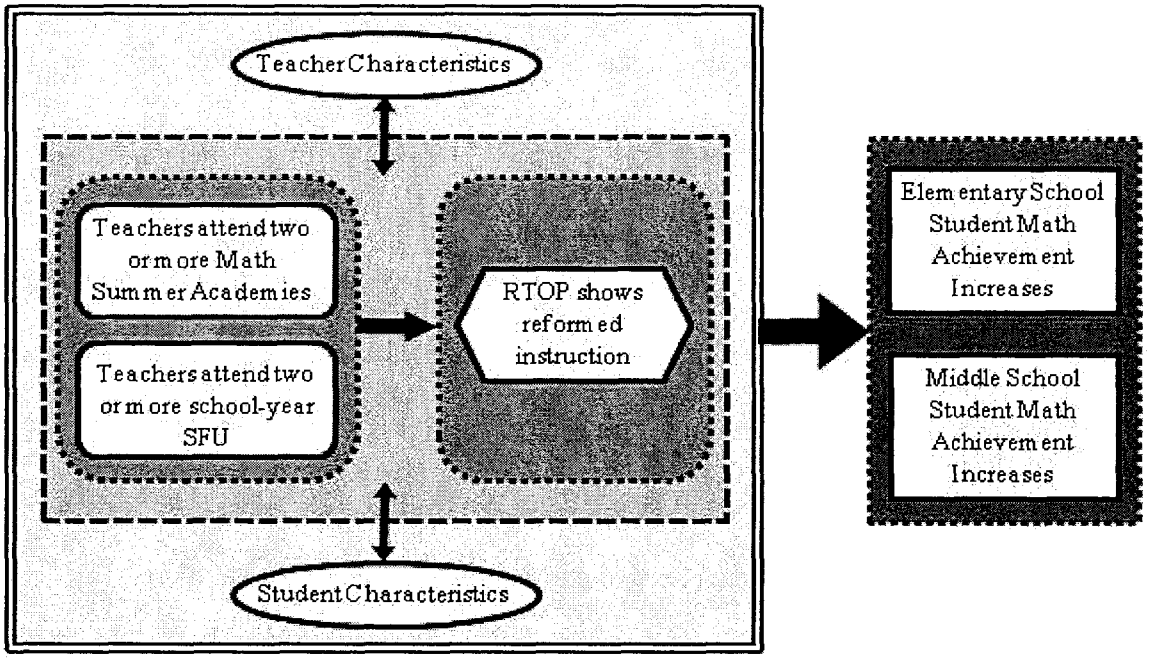


Figure 1. Flow chart expressing the theoretical linkages between the RM-MSMSP professional development, teacher learning, covariates, and student achievement.

Assumptions and Limitations

There are several potential limitations to this study both in intervention effects and in method (i.e., the ability to use this design in another context). First, this study assumes that teachers in different years and courses received a similar level of treatment.

In other words, a participant in the 2005 Algebra II course should be considered equally as treated as a teacher in the 2007 Geometry course. Second, the evaluation design is predicated on the assumption that teachers implement what they learn from the professional development in a consistent manner throughout the years used in the analysis. While the study does employ some fidelity measures, this assumption is difficult to test without observing all teacher participants at all points in time. Third, because the pre-treatment and post-treatment data (as described in Chapter 3) came from multiple years, the study results assume that historical and outside influences on teacher practice and student achievement were minimal and randomly distributed. This is possibly the biggest assumption made in this study, as there was no way to measure these potential effects.

In addition, because there was no longitudinal analysis in this study, embedded in the design is the assumption that teachers learned content and reformed teaching practices from the professional development, implemented what they learned in their classrooms, and affected student learning in one year or less, the time between treatment completion and state testing. Therefore, it is very possible that any non-significant effects were due to the lack of time given between analysis year and treatment completion.

The most salient limitation to finding intervention effects is the use of Colorado Student Assessment Program (CSAP) test scores to assert a causal link between the professional development intervention and student achievement. It is not entirely certain that the CSAP test is a good measure of an inquiry-based professional development intervention. Therefore, the outcome measures used for comparison in this study may be somewhat context-specific. However, this would only affect the ability to detect

intervention effects, and should not harm the quality of the comparisons made in the study.

Delimitations

The data used for the design comparisons made in this study have several boundaries within which the results should be confined. These boundaries should be kept in mind while reading the results, as they affect the comparisons of the two designs, although the use of the evaluation designs are not affected by these restrictions. The location of the participants in the present study provides a geographical boundary within urban schools. The partner districts were all in the Denver-Metro area, and although the school and district sizes varied, they should be considered as urban districts. In addition, this study only analyzes student data from grades 4 through 8, applying a grade-level boundary as well. These delimitations have implications to the generalizability of the study. No rural or suburban districts were analyzed, and grades K-3 and 9-12 were not analyzed in this study. The grade level boundaries were based on the goals of the RM-MSMSP program and access to data.

The teacher participants were all volunteers to the program. This leads to the delimitation that the results of either the original evaluation design or the variation design may generalize to specific types of teachers, such as those who are highly-motivated or need assistance to achieve highly-qualified status. Further, teachers who complete the program may have different characteristics than those who do not. It is difficult to determine if similar results would come from teachers of different characteristics, especially when it comes to motivation.

The outcome measures all came from a high stakes test at one point in time each year. As stated previously, there was no longitudinal component to this study. Not only does this affect the assumptions made here, but it affects the ability to draw conclusions about what happens in classes over time. Neither design accounts for multiple measures that would show if effects persist or do not persist over time. In addition, because the outcome measure is a high stakes test, there is no way to know what the results of the study would be had a different outcome measure been used.

CHAPTER 2 – REVIEW OF THE LITERATURE

This study incorporates theory and past research in a variety of domains. The main goal of this comparative study is to inform the theory on a new methodological approach. Therefore, the first section is situated current and past evaluation designs that have been used to evaluate a variety of professional development programs. Given the lack of research and evaluation studies directly tied to professional development and student achievement, especially in the middle grades, the second section is situated around research in the areas of teacher impact on student achievement. The final section gives an overview of the professional development structures that are considered to be the best at improving teacher learning, teacher practice, and student achievement.

Evaluating Professional Development

Given the focus on the quality of professional development, it is important to be able to evaluate the effectiveness of different programs. Guskey (2000) gives four reasons why evaluation is important to professional development and education reform. First, educators have become increasingly aware of benefits of the ongoing long-term process of professional development compared to the one-shot professional development events. Second, professional development is a systemic process requiring that information on program goals be gathered, analyzed, and meaningfully reported. Third, there is a need for better information to guide reforms in professional development. Fourth, there has been an increase in the focus on accountability at all levels of education, and programs must show that they are important, meaningful, and productive. So how,

then, are professional development programs being evaluated and what seem to be the best methods?

Evaluation Designs

While the benefits of professional development are known and the call for thorough evaluations of professional development programs is prevalent, Muijs and Lindsay (2008) assert that evaluations are rarely undertaken in a methodical and rigorous manner. Guskey (2000) claims that several mistakes are being made in evaluation practices in relation to professional development. First, he states that some "evaluations" of professional development are not evaluations at all, but are summaries of the activities offered as part of the program. These "evaluations" serve solely to document items like course attendance, credits accrued, and descriptions of topics presented. Although this type of documentation is useful in the context of a broader evaluation, on its own, it offers little information on the effectiveness of the program. Next, he states that most evaluations consisting of more than merely documentation are shallow, being concerned, for the most part, with teacher satisfaction. Although teacher satisfaction with the program gives insight to program aspects successful in keeping teachers involved, it does nothing toward understanding how the program effects gains in knowledge and practice. Finally, Guskey (2000) states that professional development evaluation initiatives are usually too brief to capture long-term change. Instead, he suggests that evaluation efforts be long-term and occur alongside professional development activities.

Muijs and Lindsay (2008) conducted a survey-based study on a randomly selected sample of teachers and professional development coordinators in England to understand the extent to which they felt professional development activities had been evaluated and

how they were evaluated. Based on the evaluation hierarchies of Guskey (2000), the survey consisted of questions concerning the following levels of evaluation: participant satisfaction, participant learning, organizational change, participant's use of learned knowledge and skills, student learning, and cost-benefit. Participant satisfaction was the lowest level of evaluation while student learning was the highest. Muijs and Lindsay (2008) determined that participant satisfaction was the most frequently evaluated aspect of professional development with 75% of professional development coordinators and 85% of teachers responding that it is usually or always evaluated. Teacher learning was also frequently evaluated according to both professional development coordinators and teachers (43% and 58%, respectively). All other aspects of professional development received 30% or less of "usually" or "always" responses from both coordinators and teachers. Therefore, if the self-reported data are to be trusted, there does seem to be the use of multiple levels of evaluation being employed in English schools. However, the use of the higher-level evaluation methods is still somewhat limited. In addition, evaluations that use higher-level methods also seem to use more sophisticated methods of measurements (i.e., interviews compared to questionnaires) (Muijs & Lindsay, 2008).

Quantitative methods. Randomized experimental methods seldom seem to be employed in professional development evaluation designs, which is consistent with their employment in the field of educational research in general (Porter, Blank, Smithson, & Osthoff, 2005). Even with the use of a true experimental design, there are still several weaknesses that appear in the literature. First, it is difficult to separate professional development effects of one program from another, because teachers, even if randomly selected and assigned to groups, will usually participate in more than one program. It is

not ethical or possible to conduct these experiments in a laboratory setting, so controlling for these outside factors is difficult. Even with random assignment, the participation in multiple programs cannot be assumed to be random in nature (Porter, et al., 2005). Second, when evaluating programs that reach several schools or districts, there may be unequal attrition between the schools/districts. Third, teachers might move from one school/district to another, which could mean that experimental teachers move into control schools/districts or vice versa. All of these weaknesses would impact the internal validity of the studies in ways that might not impact other true experiments.

Because experimental research in education is impractical, as was just shown, quasi-experimental research designs are often employed, as they still allow a method for inferring causal effects (Maxwell, 2004; Shadish, et al., 2002). Causal inference from quasi-experiments, as for randomized experiments, must meet three basic criteria: (1) the cause must precede the effect, (2) a relationship must exist between the cause and effect, (3) and there should be no other plausible explanation for the effect (Marini & Singer, 1988; Shadish, et al., 2002). The first two requirements can be easily achieved in both experiments and quasi-experiments by first manipulating the treatment to force its occurrence before the outcome and then assessing covariation between cause and effect using statistical analysis. Experiments address the third requirement with random assignment, which ensures that other plausible explanations are arbitrarily distributed over the experimental conditions (Shadish, et al., 2002). Because quasi-experiments do not use random assignment it is necessary to use other principles to rule out alternative explanations. Shadish, et al. (2002) emphasize three principles to address this

requirement: (1) identification and examination of reasonable threats to internal validity, (2) design-implemented control, (3) and complex pattern matching.

The first principle is fairly straightforward; Reichardt explains that once the plausible alternative explanations are identified, the likelihood that they explain the causal relationship can be analyzed (as cited in Shadish, et al., 2002). The second principle is accomplished in quasi-experimentation by preventing the confounding of threats to validity through the addition of design elements and providing evidence on the plausibility of those threats. While Shadish, et al. (2002) note that design elements and statistics can be used together, they recommend the use of as many design controls as possible and only limited use of statistical controls. They also stress that even adhering to these principles does not provide the rationale for causal inference that a randomized experiment does. Instead, implying causal inference in quasi-experiments requires that detailed attention be paid to finding and mitigating the plausibility of other explanations. The third principle builds on this necessity by requiring complex hypotheses to be made, which inherently reduces the existence of plausible alternative explanations.

While quasi-experimentation can be used as a relative substitute for true experimentation, not all quasi-experimental designs are as credible as others. In fact, some designs lead to many issues with internal validity, which makes ruling out alternative hypotheses difficult. The one group posttest-only design (Shadish, et al., 2002) is quite often used in professional development evaluation, usually when group matching and pre-program assessments cannot be done (Huziak-Clark, et al., 2007). Because of the lack of a control group and no pretest measure of what would have occurred had the professional development not been present, Guskey (2000) mentions

that with this design, these studies usually have little more than descriptive statistics in their findings. Even with the internal validity issues, most of the inferential analyses based on this design employ regression, hierarchical linear modeling, or other statistical inference techniques to evaluate professional development program effects (Huffman, et al., 2003; Huziak-Clark, et al., 2007; Supovitz & Turner, 2000).

Adding a pretest measure to a design with no control group gives "weak information about the counterfactual inference concerning what might have happened to participants had the treatment not occurred," (Shadish, et al., 2002, p. 108). Studies that utilize this evaluation design might test or survey teachers before and after participating in the professional development (Basista & Matthews, 2002). While this is an improvement over the one group posttest-only design, Shadish, et al. (2002) note that because the pretest measure takes place at a different point in time than the posttest measure, there could be alternative explanations that have nothing to do with the professional development, such as maturation or history.

Adding a control group controls for the internal validity problems associated with the previous design by offering more information about what happens with a different group measured at the same time as the experimental group. Because they are measured at the same time, maturation and history effects should be controlled (Shadish, et al., 2002). This design appears to be the most common in professional development evaluations when having a control group is possible. While this design is an improvement over the other two designs, introducing a control group brings with it other threats to validity such as selection bias. However, Shadish, et al. (2002) note that using the control group in combination with pretest measures allows for the exploration of the size and

direction of any selection bias that may exist. If pretest measures are not available and a mixed methods approach is used, there might be more confidence in the results due to triangulation (Radford, 1998).

Qualitative methods. Evaluations that are more concerned with understanding participants' perceptions, interactions, and experiences than implying a cause and effect relationship would use qualitative evaluation designs. Grounded theory, which Charmaz (2006) states is suggestive, inconclusive, and incomplete, seems to be the most common approach to qualitative evaluation. Using coding techniques, researchers uncover themes that lead to conceptual groups reflecting shared traits between participants (Panizzon & Pegg, 2008). Grounded theory-based evaluations utilize interviews, videos, classroom observations, and document analysis to draw on the nature of program.

Two other less-utilized qualitative methods are vignette analysis and case study analysis. Vignette analysis is often the next step after coding based on grounded theory. The purpose of using the vignettes in combination with the coding is to illustrate specific features of the interactions while preserving the complexity of the context in which they took place (Borko, et al., 2008). Vignettes are not transcriptions of the interactions, but rather representations of the events, people, and activities that take place. While coding can help understand themes, vignette analysis offers insight into moods and feelings, which give an overall picture of the atmosphere of the professional development.

As with vignette analysis, case study evaluations provide insight into the culture of a system rather than an issue or problem (Creswell, 2007). However, unlike vignette analysis, case study analysis provides insight on only a few participants. Participants are often chosen for how they have benefited or not benefited from professional development

(Ross & Bruce, 2008). In other words, cases in case study evaluations are often participants that have been affected in the extremes. Case study research is often used to *describe* a cause-effect relationship between the professional development and the outcome of interest, as it limits the sample size to only a few individuals (Creswell, 2007). Because of its usefulness in describing professional development program outcomes and processes, qualitative analysis is often employed in conjunction with quantitative analysis in an overall evaluation design.

Measuring Student Achievement

The majority of the studies in the literature surrounding the assessment of student achievement related to teacher characteristics and professional development have based their findings on testing measures. The measures were mainly derived from standardized tests that contained multiple-choice and/or open-ended problem-solving questions (for example, Clotfelter, et al., 2007; Hamilton, et al., 2003; Hill, et al., 2005; McCaffrey, et al., 2001). According to Popham (1999), standardized testing implies that a test is "administered and scored in a predetermined, standard manner" (1999, p. 8). Depending on the tests and where they are administered, these tests might be compared across schools, districts, or states, because they are given under similar conditions (NCLB, 2001). While many standardized tests are based on multiple choices questions, they can also incorporate open-ended responses and essay-type questions.

Standardized testing can take one of two forms in terms how the outcomes affect the student: high-stakes or low-stakes. High-stakes testing involves consequences that are attached to the results, which can affect the school, the student, or both. State testing is commonly seen as high-stakes, because of the necessity to maintain Adequate Yearly

Progress (AYP) as defined by NCLB (2001). Some states also attach a high-stakes component to students by tying their graduation or grade-level promotion to the results of the test (Cizek, 2001; Resnick, 2004). In contrast, low-stakes testing has no consequences outside of the classroom (i.e., there may be a classroom grade attached). Low-stakes tests might help the teacher better understand how students are learning in order to alter classroom practice.

While standardized tests are designed for comparison, using them to evaluate the effects of teachers or programs on student achievement comes with several caveats. Goe (2007) states that standardized tests "were not engineered to be particularly sensitive to small variations in instruction or to sort out teacher contributions to student learning from other factors..." (p. 15). Student learning may grow at the desired rate, but tests that are not aligned to state standards, school curriculum, or classroom practices may not measure this growth. Although these tests are not suitable for drawing direct conclusions about teachers, using standardized test scores for research under certain conditions to inform a knowledge-base may be useful if there are not other comparable outcome measures (Goe, 2007).

When discussing student achievement, it is often necessary to discuss how students change over time. One method for analyzing student change is a gain score. In its simplest form, a gain score is just the difference between two measures, one taken before the other (Linn, 1981). However, simple gain scores can be unreliable if the pretest and posttest are highly correlated, which is often the case. Still, Linn (1981) notes that while the reliability problem is a concern when used to make decisions about individuals, it becomes less of a concern when analyzing groups.

Gain scores are also usually correlated with the pretest, which is a concern, because the "implicit goal in the use of [gain] scores is often to remove or adjust for initial differences and thereby make it possible to compare the gains of individuals or groups that started with unequal pretest scores" (Linn, 1981, p. 87). The consequence of correlated gain scores and pretest measures is that students who have a lower pretest score will show greater gains than those who have higher pretest scores. Therefore, in studies interested in analyzing the differences between two groups of students there will be built-in bias toward the group that initially scores lower on the pretest.

Although the reliability of gain scores has been questioned for years, Zimmerman and Williams (1998) give situations where the reliability can actually be quite high. If the variance and reliability of the posttest are larger than that of the pretest, the reliability of the gain score can be sizeable. The same is true if the variance and reliability of the pretest are larger than that of the posttest. These conditions can be met in a situation where an intervention increases the variance of the posttest measure. In fact, Zimmerman and Williams (1998) assert that gain scores will only be unreliable if the intervention decreases the variance in posttest scores, and gain scores can be reliable for research if the pretest measure is reliable. Therefore, if a reliable pre-test measure is found, it is reasonable to use gain scores as a means to measure student change from a pretest to a posttest.

Teacher Impact on Student Achievement

Although teachers do not ultimately cause student learning, they can positively affect student learning by providing students with time, a classroom climate conducive to learning, structure, and meaningful activities that support learning (Lasley, Siedentop, &

Yinger, 2006). Teachers who are more highly qualified, as defined by NCLB, are thought to have a greater impact on student learning than those who are not (Hanushek, 2006; Rockoff, 2003). In fact, research during the last decade has examined the relationship between teacher quality and student achievement and provides evidence that quality teaching does make a difference in student learning, because teacher expertise has been shown to be a significant variable in the function of student achievement (Ball, Lubienski, & Mewborn, 2001; Laczko-Kerr & Berliner, 2002; Lasley, et al., 2006). However, what characteristics make a teacher high quality and what determines an effective teacher?

Content Knowledge

Subject matter knowledge is a variable that is often assumed to affect student achievement. The literature suggests that perhaps this relationship is not quite so straightforward. Much of the literature surrounding teacher content knowledge impacts on student achievement focus on using proxies, such as teacher education level, academic ability, degrees, and certifications. "By using such measures, researchers implicitly assumed a connection between formal schooling and...aspects of teachers' knowledge and performance that produce student outcomes," (Hill, et al., 2005, p. 374). Darling-Hammond (1999), in her synthesis on teacher effects on student achievement, showed that student achievement is impacted by whether or not teachers are certified in mathematics. Further, she states that certified teachers who also have had pedagogical training and have pedagogical content knowledge have a greater impact on students. Goldhaber and Brewer (2000) backed up this claim in their study by showing strong evidence that students who had classes with teachers certified in mathematics

outperformed those whose teachers were not certified. While the literature provides further evidence that teacher certification does impact student achievement (Darling-Hammond, Berry, & Thoreson, 2001; Wilson, Floden, & Ferrini-Mundy, 2001), Rivkin, Hanushek, and Kain (2005) found that there is minimal variation in student gains from differences in teacher qualifications such as certification. Therefore, it appears that the findings on the relationship between teacher certification and student achievement are mixed.

Similarly, in a review of the literature on teacher characteristics related to student gains, Wayne and Youngs (2003) concluded that, overall, the relationships between teacher degrees and coursework and student gains are ambiguous. After disaggregating studies by subject area, they found that there was a positive relationship between the number of college mathematics courses taken by teachers and their students' gains at the high school level. However, there is a lack of research surrounding elementary and middle school student achievement related to teachers' coursework. In a multilevel analysis of the Longitudinal Study of American Youth, Monk and King (1994) found both positive and negative effects of teacher subject matter preparation on student achievement, while Monk (1994) found that these data created a curvilinear trend. This suggests that perhaps teacher content knowledge, as derived from the number of courses taken, may have a threshold effect where past a certain number of courses, the effects diminish. While the results of these studies could be conflicting because of methodological issues (Greenwald, Hedges, & Laine, 1996; Hanushek, 1981, 1996), a more salient cause could come from the possibility that these measures are simply not good proxies for teacher content knowledge.

Rowan, et al. (1997) used data from the National Education Longitudinal Study of 1988 to analyze 10th grade student mathematics achievement related to teacher content knowledge as measured by a math quiz. Results showed that students of teachers who answered the math quiz correctly had higher achievement levels than students of teachers who answered incorrectly. In a study of 89 schools, Hill, et al. (2005) found that teacher responses to researcher-constructed mathematical content problems significantly predicted first and third grade student gains in mathematics. Although these two studies that use actual measures of teacher content knowledge show positive relationships between teacher knowledge and student achievement, there is some question concerning the lack of alignment between the teacher content knowledge and student achievement measures (Hill, et al., 2005). Further, these two studies look at early elementary school and high school, but there have been no middle school studies that directly measure teacher content knowledge.

Instructional Practices

The NSF and other groups promote the use of reform-based instructional practices, which hinge on engaging students as active participants in their learning. Reformed teaching emphasizes problem solving, communication, reasoning, and mathematical connections at every grade level and within content standards (NCTM, 2000). Specifically, reform teachers (a) view classrooms as communities of learning; (b) encourage the use of logic and evidence to draw conclusions; (c) promote mathematical reasoning in place of memorization; (d) provide connections between concepts and applications; and (e) focus on inference, discovering, and problem solving (NCTM, 1991).

The reformed classroom described here fits the NSF (2000) definition of "inquiry" as "a process of exploring the natural or material world...that leads to asking questions, making discoveries, and rigorously testing those discoveries in the search for new understanding" (p. 2). Inquiry, as defined here, gives children multiple ways to communicate their thoughts and ideas, allowing teachers to have direct and accurate knowledge of the inquiry process and student learning (Dyasi, 2000). Ash (2000) provides a set of process skills through which one goes in learning through inquiry: observing, questioning, hypothesizing, predicting, investigating, interpreting, and communicating. In addition to inquiry-based learning, this process of learning is also often used under the terminology "standards-based" and "reform-based" learning.

Research provides some confirmation of the effectiveness of some individual practices that fall under the blanket definition of inquiry. For example, a study undertaken by Ginsburg-Block and Fantuzzo (1998) found that assigning low-achieving elementary students to problem-solving or collaboration conditions resulted in a higher rate of correctly completed math problems compared with students who were not in these conditions. Similarly, Cohen and Hill (2000) found that teacher-reported frequency of use of reform-based classroom practices was positively related to students' scores on the state mathematics test. Schoenfeld (2002) states that students in reform-based classrooms consistently outperform students in traditional classrooms on tests of conceptual understanding and problem solving.

Experience

It is a common belief that the longer people perform a job the more proficient they become at effectively completing work. The same assumption can be made for

teachers in that the more years of experience teachers have the more effective they become at teaching and raising student achievement. There is evidence, though not statistically significant, that this assumption is true (Fetler, 2001; Laczko-Kerr & Berliner, 2002; Rowan, 2002; Wayne & Youngs, 2003).

While there is evidence that, in general, student gains increase with teacher experience, there seems to be a more complex relationship. Darling-Hammond (1999) stated, "While many studies have established that inexperienced teachers (those with less than three years of experience) are typically less effective than more senior teachers, the benefits of experience appear to level off after about five years..." (p. 9). Similar results were found in an analysis of Texas student achievement data:

> The results for teacher experience generally support the notion that beginning teachers and to a lesser extent second and third year teachers in mathematics perform significantly worse than more experienced teachers. There may be some additional gains to experience in the subsequent year or two, but the estimated benefits are small and not statistically significant in both mathematics and reading... (Rivkin, et al., 2005, p. 447).

Consistent with these findings, Boyd, Lankford, Loeb, Rockoff, and Wyckoff (2007) correlated middle level math teacher experience with student achievement in New York City Schools. The results showed that teachers tended to improve student achievement through the first three to five years of experience with the highest improvement occurring between the first and second years.

Although these previous three studies suggest a nonlinear relationship between years of experience and student achievement (i.e., there is growth mostly in the early years that flattens off in the later years), Clotfelter, Ladd, and Vigdor (2007) found that student achievement in math tended to be relatively linearly related to years of experience. In other words, teachers with more years of experience tended to be more

effective in raising student achievement. They add, however, that the largest teacher experience benefit to student achievement occurred in the first two years of teaching.

Although the literature tends to show a positive relationship between teacher experience and student achievement, in their meta-analysis of the connection between teacher characteristics and student achievement, Wayne and Youngs (2003) assert that "studies that use convincing research designs simply do not exist or have not been conclusive" (p. 107). They also noted that numerous effects are captured in the teacher experience variable (e.g., motivation, hiring timing, and differences in those who leave the profession and those who stay), but these effects make the relationship between teacher experience and student achievement too difficult to interpret.

Best Practices in Professional Development

Guskey (2000) defines professional development as "those processes and activities designed to enhance the professional knowledge, skills, and attitudes of educators so that they might, in turn, improve the learning of students" (p. 16). Traditionally, professional development has been confined to the narrow scope of short, one-shot, stand-alone workshops aimed at the general knowledge of the teacher (Ball & Cohen, 1999; Birman, et al., 2007; Guskey, 2000). These workshops are usually chosen by school or district officials, are rarely planned with input from teachers, and are only sometimes relevant to their subject areas.

Some teachers also view graduate courses as fitting into the realm of professional development (Guskey, 2000). This view fits nicely with policies that require teachers to earn a particular number of credits or hours in a year. While these policies were designed for the purpose of highlighting the importance of ongoing learning, they often lead to the

attitude of getting hours instead of focusing on what needs improvement (McDiarmid, David, Kannapel, Corcoran, & Coe, 1997). Guskey (2000) notes that it is important to create a school culture of including learning in everyday tasks rather than placing emphasis on certain days throughout the school year. Otherwise, he states, teachers will view professional development as a task that they must complete instead of a means for self-improvement.

While the one-shot, short term professional development activities are still in use, the idea of what professional development should look like began forming during the 1990s when experts began to suggest that the traditional programs were inadequate for changing teacher practice and helping teachers meet the needs of their students (Guskey, 2000, 2003a; Kennedy, 1998; Little, 1993; Loucks-Horsley, Love, Stiles, Mundry, & Hewson, 2003). What emerged were several lists of good research-based characteristics of professional development that overlapped on some points (Guskey, 2003b). Of the 21 characteristics distinguished by Guskey (2003b), there were eight common elements that appeared in each list (Garet, et al., 2001; Guskey, 2000; Hawley & Valli, 2001; Loucks-Horsley & Matsumoto, 1999):

1. Includes teacher needs and teachers developing professional development opportunities,
2. School-based and incorporated into school procedures,
3. Part of widespread transformation (i.e., systemic change),
4. Collaborative in nature,
5. Allow teachers to develop a theoretical understanding of knowledge and skills they learn (e.g., how they learn),

6. Focuses on the differences between standards and goals for student learning and performance,
7. Continuous and on-going including support for further learning,
8. Integrates an evaluation of professional development effect on teacher and student outcomes.

The National Research Council (2000) created a simplified list of four principles of learning to consider in professional development design into which the above characteristics fall. Programs should be centered on learning, knowledge, assessment, and community. Being learner-centered implies that the professional development considers the needs of the learner, while being knowledge-centered places a serious focus on helping teachers become knowledgeable in the learning process. Assessment-centered programs focus on the usefulness of both self-assessment and alternative classroom assessment, and community-based programs build community within teachers and between teachers and administrators.

While there seems to be an abundance of literature expounding the *characteristics* of successful professional development, there is very little consensus on actual successful *types* of designs. Loucks-Horsley & Matsumoto (1999) built on the NRC learning principles and incorporated the eight common characteristics to form five strategies and structures that are used in professional development design: (1) immersion, (2) curriculum, (3) examining practice, (4) collaboration, and (5) mechanisms. Immersion-based programs tend to be focused on hands-on teacher learning. Often immersion programs include teachers working closely with a professional from their field or solving problems in mathematics. Curriculum-based programs provide teachers with knowledge

and materials they will actually use in the classroom. These programs might include teachers aligning curriculum to state standards or teaching a unit on a topic that is new to them. Programs centered on examining practice tend to be employed during the school-year and are focused on actual classroom practice. These may include observations, videotaping, or self-assessment. Collaboration-based programs typically center on learning-communities within schools or partnerships between teachers, mathematicians, and learning coaches. Finally, program mechanisms are related to structures through which learning occurs or how the program is given. Mechanisms could include workshops, institutes, courses, or technology (e.g., web-based formats).

Strategies and structures are typically used in combination, depending on the goals of the program (Loucks-Horsley, et al., 2003). For example, a program could use workshops and summer institutes where teachers work together to design mathematics curriculum surrounding state standards while learning content knowledge in their field. In this example, the professional development would be focused on both curriculum and collaboration using two different delivery mechanisms (i.e., workshops and summer institutes). In addition to the complexities brought about by the intertwining of structures and strategies, an added level of complexity is brought about when there is more than one outcome, which is usually the case.

The goals or outcomes of professional development programs are more straight-forward than the types of professional development design. Loucks-Horsley, et al. (2003) classifies these outcomes into four different domains: (1) student learning, (2) teacher learning, (3) teacher practice, and (4) organization. While student learning seems to be the goal of more recent professional development programs (Huffman, Thomas, &

Lawrenz, 2003; Panizzon & Pegg, 2008), past literature seems to focus very little on this outcome. Because the relationship between professional development and student achievement is complex, it can be difficult to evaluate. So, this gap is not surprising (Borko, 2004; Guskey & Sparks, 2002).

Kennedy's (1998) seminal literature review that focused on mathematics and science professional development and student learning found strong effects ($r > .40$) of programs focused on subject matter knowledge and student learning on student performance in reasoning and problem solving. A similar review undertaken by Yoon, Duncan, Lee, Scarloss, and Shapley (2007) showed strong effects as well (average $r = .55$). Other studies also have shown positive effects, albeit at a much smaller level than was found in the reviews, of specific professional development programs on student achievement (Meyer & Sutton, 2008; Panizzon & Pegg, 2008). However, none of these studies were conducted on a large number of teachers in different settings (Wayne, Yoon, Zhu, Cronen, & Garet, 2008).

Huffman, et al. (2003) focused on the relationship between different types and duration of professional development, teacher practice, and student achievement in mathematics. They surveyed 104 middle school math teachers from 46 different schools concerning their own practice and types of professional development in which they had participated. Mean state mathematics achievement scores were used to determine student achievement outcomes, and regression analyses were done to determine which, if any, types of professional development could predict student achievement. Only one component of professional development (curriculum development) predicted student achievement, although the relationship was negative. The authors posit that these results

could relate to the previous achievement level of teachers' students. In other words, teachers who participate in long-term professional development based on curriculum development could be doing so, because their students are low-achieving, while teachers with highly achieving students have less motivation to participate. They also note that curriculum development accounted for only 16% of the change in student achievement, implying that there are other factors that were not controlled.

Most of the literature on professional development centers on teacher learning and practice (Basista & Matthews, 2002; Desimone, Porter, Garet, Yoon, & Birman, 2002; Kimmel, Deek, Farrell, & O'Shea, 1999; Shotsberger, 1999). Although these studies acknowledge that student learning is the ultimate goal of affecting teacher knowledge and practice, they do not directly measure the effects on student learning. Instead, the focus is placed on adding to teacher content knowledge as well as their pedagogical content knowledge. These programs place importance on teaching the teachers how students will learn, and at the same time, the teachers gain content knowledge. They are almost always coupled with evaluations of the change in teacher classroom practice. The literature shows that teachers who participate in sustained professional development based on structures described previously have increased subject matter knowledge, greater efficacy, and pedagogical content knowledge (Basista & Matthews, 2002; Borko, Jacobs, Eiteljorg, & Pittman, 2008; Huziak-Clark, Van Hook, Nurnberger-Haag, & Ballone-Duran, 2007; Shotsberger, 1999; Swackhamer, Koellner, Basile, & Kimbrough, in press).

Finally, organizational outcomes are more concerned with program sustainability and developing structures that would allow professional development to continue. Loucks-Horsley & Matsumoto (1999) explain that these structures should include school

and district administrative support and create learning communities that allow teachers continued support in their learning and practice. For example, a program may utilize common teacher planning periods that could continue long after funding for the actual program runs out if the proper structure is in place and a culture of learning has permeated through the school. Organizational goals may exist as one of the major goals of the program or they may be secondary to the other program goals (Loucks-Horsley, et al., 2003).

While some professional development programs focus on one outcome, most tend to incorporate several outcomes. For example, Panizzon & Pegg (2008) were interested in studying if their SOLO model of professional development would be successful in both changing teacher practice in relation to assessment use in the classroom and enhancing student mathematics and science learning. Similarly, the RM-MSMSP program was interested in all four outcomes by creating a sustainable program that would act to increase teacher content knowledge and change teacher practice, which, in turn, would increase student mathematics achievement. Incorporating all four outcomes into professional development program goals is not uncommon, because they are so closely tied together (Loucks-Horsley & Matsumoto, 1999).

Overall, research shows that a context of teachers reflecting, accessing new ideas, experimenting in the classroom, and sharing experiences with other teachers within schools where teachers also receive the support of administrators, schools, and classrooms has a great potential for improvement (Muijs & Reynolds, 2000). In addition, giving attention to teacher learning can have direct and indirect impacts on student attitudes toward learning, beliefs about their own learning potentials, and achievement

(Joyce, Calhoun, & Hopkins, 1998). However, the National Research Council (2000) concluded that "even when resources are formally provided for teachers' continued development, opportunities for effective learning vary in terms of quality," (p. 192).

CHAPTER 3 – RESEARCH DESIGN AND METHODS

The inherent complexity of the professional development in this study made choosing a sound research method difficult. Because of the lack of a teacher control group and non-randomization of both teachers and students, a quasi-experimental design was both necessary and problematic. Shadish, et al. (2002) point out that quasi-experiments, given their non-random nature, usually create less compelling arguments for causation. For example, in the RM-MSMSP program, teachers elected to participate and could discontinue participation at any time. Therefore, a spurious relationship might exist between the professional development and student achievement simply because those teachers who completed the treatment as defined by program implementers were more motivated than those who discontinued participation. Still, information on causal mechanisms is important for program implementers to understand how their program is performing in order for them to strengthen the successful components and remove unsuccessful ones. Funding organizations expect program implementers to show that their program works in the manner in which it was proposed. Further, if program theory is not shown to be true, some kind of research-based explanation is expected. Although ideal, the random selection of participants and their assignment to groups are often not feasible for large professional development programs like the RM-MSMSP. To evaluate the student achievement outcomes of the RM-MSMSP project, a variation on the quasi-experimental cohort control design (Shadish, et al., 2002) was conceptualized. This study compares this variation design with the original cohort control design to test the variation

design's utility for the evaluation of the RM-MSMSP program and similar professional development programs.

This chapter first discusses the design on which the variation was created and subsequently compared. Using the information about the RM-MSMSP project, the evaluation is described in terms of the original cohort control design and then the variation in order to explain the subtle, but important, differences between the two designs. Next, a thorough description of the intervention, participants, and procedures for the quantitative data used in the comparisons are provided. Then, threats to internal validity as listed by Shadish, et al. (2002, pp. 56-61) are discussed in a general sense. The variables used for the final analysis are then presented along with the methods for data collection and management. Finally, a brief overview of the quantitative analytical procedures used to address the stated research questions is given. Note that a more detailed discussion of threats to internal validity can be found in Chapter 4, as internal validity was used as a measure to compare the two evaluation designs. In addition, the external reliability is discussed in Chapter 4 in terms of the utility of the two evaluation designs compared in this study.

The Cohort Control Design

There are many quasi-experimental designs from which to choose. Although frequently used, many of these designs provide a weak basis for causal inference, because they lack a control group, pretest, or both. Shadish, et al. (2002) assert that designs using a *carefully selected* control group can assist causal inference, but are better used in combination with pretests. These pretests help explain how the groups initially differ, allowing for a stronger assumption that the groups are actually similar.

Shadish, et al. (2002) describe what they call the "cohort control" design as useful for institutions and programs that experience regular turnover or movement from one year to the next. The cohort control design utilizes a control group, but no pretest measure. This means that there is information on a group that did not participate in the treatment and information on a separate group that did participate in the treatment. However, there is no information on the outcomes at more than one point in time. Also, the control group is measured at a different point in time (before the treatment took place) than the treated group (after treatment took place). This has internal validity implications that are discussed later in this chapter. The cohort control design can be depicted as:

NR	O_1		
NR		X	O_2

where NR means both the control group and the treated group were non-randomly selected, O_1 is the pre-treatment measure on the control cohort , X is the treatment, and O_2 is the post-treatment measure on the treated cohort. A "cohort" refers to successive groups in which one group follows the other, such as students moving up a grade-level in school. In this example, Cohort 1 could be sixth grade students in 2005 while Cohort 2 could be sixth grade students in 2007. The students remain in their respective cohorts throughout their time in school until they graduate or change schools. "The crucial assumption with cohorts is that selection differences are smaller between cohorts than would be the case between non-cohort comparison groups" (Shadish, et al., 2002, p. 149). The researcher must investigate this assumption by examining background characteristics believed to be connected to outcomes, such as, in the student achievement case, ethnicity

or gender. In such cases, the control and treated groups should be compared to reduce the risk of alternate explanations being present.

The cohort control design can be particularly useful in studies with the following four characteristics:

1. One cohort experiences a given treatment and earlier or later cohorts do not;
2. Cohorts differ in only minor ways from their contiguous cohorts;
3. Organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and
4. An organization's archival records can be used for constructing and then comparing cohorts. (Shadish, et al., 2002, pp. 148-149)

The cohort control design seems to be mostly used in the medical sector. One example comes from a retrospective study of the merit of aspirin therapy on pregnant women who were not fully diagnosed as having Antiphospholipid Syndrome (APS). The researchers used data from one hospital's patients who had previously miscarried, but had not been diagnosed with APS. All of the patients had received the same treatment except for the administration of aspirin to the experimental group, a practice that is common for the prevention of pregnancy loss in APS patients. Although, this research design is typically employed in medical research, one education-based study was conducted by Minton (1975) who analyzed the effects of the first year of *Sesame Street* on Kindergarten student performance on the Metropolitan Readiness Test. To do this, she compared the scores from a cohort of older siblings who had not been exposed to the television show as the control group to a cohort of younger siblings who took the test after being exposed to the program.

The RM-MSMSP study on student achievement used a cohort of students from the year before teachers began participating in the program as a control and a separate cohort of students from the year directly after teachers became treated in the program as the experimental group. Teachers were considered fully treated after they took at least two courses and two SFUs. Therefore, using students from teachers' classes before participation in RM-MSMSP was equivalent to a control group, because this group did not benefit from the treatment while students in the classes after treatment did. For the second characteristic presented by Shadish, et al. (2002) to be true, the two cohorts were analyzed for differences in demographics and history. Although partner districts did not insist that everybody be treated, it was assumed that all students of treated teachers had the opportunity to benefit from the treatment after teachers participated in the program. Finally, the partner districts made their archival records available, so the study met the fourth criteria.

The following two sections illustrate the RM-MSMSP evaluation in terms of both the cohort control design and a variation on the design. The first section describes how the evaluation was done from the perspective of the original cohort control design. The second section describes the evaluation in terms of the variation of the cohort control design.

Evaluation with the Cohort Control Design

The outcome measure in this study was student achievement based on mathematics scores on a state high-stakes test. Student lists for both pre-treatment and post-treatment cohorts were used to identify students in each teacher's classes resulting in

a set of pre-treatment scores and post-treatment scores. The groups of scores were made from different groups of students who had the same teacher in different years.

Because the teachers in this dataset all began their participation in the summer of 2005, the pre-treatment data came from the spring of that year, giving a measure for participant teachers' students directly prior to their participation in the RM-MSMSP program. Also, records on when each teacher became treated were provided by the partner districts, allowing us to know the post-treatment year. Teachers did not all enter the program or become treated at the same time, so, practically speaking, there could be more than one set of pre-treatment students who make up the overall pre-treatment group and the same for the post-treatment group. In this comparison, however, only teachers who became treated in the spring of 2007 were selected for reasons described later in the chapter. Therefore, the pre-treatment group was made up of the Spring 2005 students in the classes of teachers who eventually became fully treated, and the fully treated teachers' students in Spring 2007 comprised the post-treatment group. It is important to note that the pre-treatment and post-treatment data came from two different years and two completely different groups of students, and there was no measure of previous performance on the tests, allowing for external factors to possibly provide an alternative explanation for the cause-effect relationship. To account for some of the external factors, it was important to examine the demographics of the students in the pre-treatment and post-treatment groups before beginning any data analysis to make sure that one group was not considerably different from the other. However, as is discussed later, this examination of the data was not sufficient to account for all possible external influences.

Evaluation with the Variation of the Cohort Control Design

Because the cohort control design does not control for students' prior performance, a major modification was made to improve causal inference. The variation of the cohort control design here utilizes a pretest of sorts for both the pre-treatment and post-treatment groups. The control teachers (pre-treatment group) and treatment teachers (post-treatment group) are the same, which might lead one to believe that the design should be purely pretest-posttest in nature. However, this is not the case, because the students in each group are different. Similarly to the cohort control design evaluation described previously, in this variation, the pre-treatment data were collected on all teacher participants using testing data on students the teachers taught prior to becoming involved in the professional development program. The post-treatment data were collected on a different group of students of the same teachers, those the teacher taught after completing the professional development sequence. Because the student groups are different, the pre-treatment data should be considered as a time-lagged control group to the post-treatment data, just as occurs in the original cohort control design.

The pre-treatment and post-treatment conditions involve the use of two separate observations to obtain a gain score. This variation on the post-test only cohort control design can be depicted as follows:

$$\frac{2NR \quad O_{1a} - O_{1b}}{2NR \quad X \quad O_{2a} - O_{2b}}$$

where 2NR reflects that both teachers and students were non-randomly selected. For this study, O_{1a} represents a posttest vector of CSAP total scaled scores for all of a participant teacher's mathematics students in the spring of the year directly before beginning

professional development (pre-treatment index year), and O_{1b} reflects the pretest vector of scaled scores for the same students in the year prior to this pre-treatment year. X reflects the professional development treatment. O_{2a} symbolizes a posttest vector of scaled scores in mathematics for the teacher's students from the CSAP test in the spring directly following the professional development treatment (post-treatment index year). O_{2b} is the pretest vector of scaled scores for the year directly prior to the post-treatment year for the same students as in the post-treatment year. An important distinction between the original cohort control design and the variation is the use of the gain score, which provides a concise way of including a pretest measure for both groups.

Because the terminology used to describe the variation design can be confusing, it is worth a reminder that the use of "pretest" and "pre-treatment" are different, and the same is true for the use of "posttest" and "post-treatment". In this design the term "pre-treatment" refers to a group of students in the spring of the year directly prior to teachers' participation in the professional development program. The term "post-treatment" refers to those students in participant teachers' classes after they became treated per the program's requirements. However, the terms "pretest" and "posttest" refer to set of scores in each group (pre-treatment and post-treatment) that were used to make up the overall gain scores. In other words, as can be seen in the previous depiction, the students in each group have a pretest and posttest score.

Shadish, et al. (2002) recommend using pretests on both the control group and treatment groups to analyze group differences before comparing them on the outcomes measure. Although using the same teacher in both pre and post-treatment groups reduces the risk of selection bias based on the teacher, it does not reduce the risk of the students

being different. Therefore, student demographic data on variables known to affect student achievement (e.g., gender, ethnicity, and special education status) were used in this design to compare pre-treatment and post-treatment student groups for each teacher.

As stated earlier, even after making sure that there are no major differences between the two groups on their demographics, the difference in time between the pre-treatment measure and the post-treatment measure could introduce bias. To account for this time bias, the design variation takes pretest scores into account through the use of gain scores. The gain scores reflect how much student scores change over time. The main assumption in this piece of the design is that students would have similar gains from year to year unless some sort of occurrence (e.g., the RM-MSMSP professional development program) takes place to make the gains change. The gain score for the pre-treatment students represents how much the participant teachers affected student performance before participating in the professional development. The post-treatment gain scores represent how much the teachers affected their students after their participation in professional development. Therefore, based on the theory that the professional development positively affects student achievement, teachers would have more of an effect on their students after they participated in the professional development. So, the difference between the pre-treatment and post-treatment gain scores represents the amount of the effect of the professional development program on student achievement.

The use of gain scores as a pretest measure is the primary difference between the variation and the original cohort control design. Using gain scores eliminates the need to run an analysis on four variables, a "pretest" and "posttest" measure for each group. While within-subjects error terms can be employed when using raw pretest scores to

compare between the groups, the gain scores incorporate the pretest into the outcome measures creating a more parsimonious design.

Professional Development Intervention

The RM-MSMSP professional development program was designed to offer math and science content courses during the summer and structured follow-up (SFU) units during the school year. The RM-MSMSP courses were designed to comprise about 80% math or science content and 20% pedagogical information and were taught in two 2-3 week summer institutes. Teachers chose to take up to two courses during the summer. The courses were followed by school year SFU courses taught across four Saturdays that were designed to focus on 20% content and 80% pedagogy. Mostly, however, the focus was on teaching content to teachers, as even the SFUs were based on the content provided in the summer institutes. As an example of the course sequencing, if Ms. Smith took Algebra I during the first summer session and Algebra II during the second summer session, she could then take the Algebra I SFU in the fall and the Algebra II SFU in the spring. After participant teachers completed this sequence of two courses and two SFUs, program leaders considered them to be "treated". However, program leaders did not institute a time limit on how long participants had to complete the sequence. In addition to the summer course and school-year SFU structure, there were courses taught during the school year that encompassed both the summer course and the SFU, allowing teachers to take a shorter amount of time (one academic semester) to complete what would usually take a summer and a school year semester.

During the first three years of the project, all courses were taught by three trained faculty members, one from the College of Liberal Arts and Sciences at the University of

Colorado Denver (UCD), one from the School of Education and Human Development at UCD, and one from the K-12 partner districts. Project teaching faculty members participated in professional development activities throughout the length of the program, which were designed to teach them how to model the concepts that they taught in the RM-MSMSP courses. Development teams met on a regular basis to develop and discuss course content, debrief after summer courses, and review course assessments.

The RM-MSMSP program paid participants a $3000 stipend for taking at least one summer course and school year SFU, and they received a prorated amount for attending only the summer institute. Each RM-MSMSP course carried university course credit (four credit hours per course), and participants were allowed to apply for up to 20 hours of credit toward a Masters degree in the School of Education and Human Development at UCD.

Sampling Design

Teachers

The RMMSMSP project relied on brochures, a website, and word of mouth to recruit participants. The demand for the program led the RM-MSMSP project staff to create a triage system based on project goals, wherein they prioritized who they accepted into the program for any given year. The triage system prioritized participants in the following manner:

1. Middle school math and science teachers who were highly qualified or seeking highly qualified status in one of the seven partner districts;
2. Middle school teachers from partner districts who did not teach math or science (e.g., special education, librarians, and other content teachers);

3. Highly qualified Elementary or high school math or science teachers from partner districts;
4. Partner district elementary or high school teachers who did not teach math or science;
5. Non-partner district teachers and faculty of any grade level and any subject.

As of the summer of 2008, 374 teachers had taken at least one math or science RM-MSMSP course. For the purposes of this study, only participant data from those teachers in any of the seven partner districts who had taken math courses, taught math in their schools, and were fully treated by Spring 2007 were used. Although there were teachers who were considered treated by Spring 2008, these data were not available at the time of the analysis, and were not included in this study.

Teachers were not assigned to courses by program staff. Rather, the teachers submitted their top three course choices each summer, and courses were filled based on the triage process described above. Teachers did not enroll in courses as cohorts, so it was possible that a teacher who had participated for two years was in the same class as a teacher who was new to the program. The program was designed to be menu-driven, allowing the teacher participants to have a choice in what courses they took. Therefore, there was no pattern in course enrollment other than the prerequisite requirements that some courses had.

In all, seven of the original 374 teachers appear in the Spring 2007 dataset analysis, which was used to compare the two evaluation designs. Several things contributed to the small number of teachers whose data were usable for this comparison. First, only teacher participants who were fully treated by Spring 2007 were considered.

Of the 82 teachers who were fully treated by Spring 2007, there were only 43 teachers who had been treated in math. It was important to look at math teachers, because this comparison study uses math test data. Students of science teachers who were fully treated in math were not included, because the math concepts being tested on the CSAP are not directly taught in science classes.

Only ten of the 43 math teachers who were fully treated in math stayed in the same district and taught math during the four years in which data were taken. Upon inspection of the numbers of students taught by these ten teachers in each grade level, it was decided to only include those teachers who taught sixth grade, because there appeared to be a tendency for teachers to teach higher grade levels in the post-treatment year than in the pre-treatment year. This could bias the outcomes, as is discussed later in this chapter, which would make comparing the two evaluation designs more difficult. It should be noted that sixth grade can be taught at both the elementary or middle school level, depending on the participant school district. However, the teachers who taught sixth grade in both the pre and post-treatment years turned out to be elementary teachers. Because this study is only concerned with the comparison of two evaluation designs and not with interpreting the outcomes of the evaluation, the final sample of teachers included the seven teachers who were fully treated in math and taught sixth grade math at the elementary level in both the pre-treatment and post-treatment years.

One of the largest problems in this study was tracking teachers for all four years, as many teachers moved to larger districts or were promoted to administrative (non-teaching) roles. In addition, some math teachers also taught science. Some years, they taught both math and science, which allowed for their inclusion of their math students'

data in the dataset. However, many teachers who initially taught math had switched to teaching science in their post-treatment year. In addition, many of the math teachers who had post-treatment math data had taught only science in their pre-treatment year.

Students

Students were found for this study based on each teacher's class lists, which were provided by the partner districts each year. All of the students in each teacher's sixth grade classes who had both a pretest and posttest score were selected as the sample. It was assumed that students were not randomly assigned to classes, and therefore were non-randomly selected in this study as a convenience sample. Pre-treatment students were selected from the teachers' course list for the spring directly prior to their involvement in the RM-MSMSP program. This study only analyzed data for students of teachers who became treated in 2007, and all of these teachers began participation during summer 2005. Therefore, the pre-treatment students came from participant teachers' classes in 2005. Similarly, post-treatment students were selected from the teachers' course lists for 2007. Table 1 summarizes the demographic make-up of the two groups of students in each year.

Only students who had both a pretest and posttest score were included in the final analysis. For example, if a pre-treatment student had a score for the spring directly before his teacher participated in professional development, but did not have a score during the year directly prior, he was not included in the analysis. A total of 265 students were used for the analysis of the 2007 data; 118 students were in the pre-treatment group and 147 were in the post-treatment group. The proportions of students in each demographic group were similar relative to the overall pre and post-treatment sample sizes. As discussed

earlier, Shadish et al. (2002) recommended the cohort control design for studies in which there are no large differences in the pre-treatment and post-treatment groups. Table 1 confirms that, at least in terms of demographic differences, the cohort control design and the variation design could be appropriately employed in this study.

Table 1

Student-level demographics in each treatment group for each data analysis year. IEP refers to students in an individualized education program.

	Pre-Treatment (N = 118)	Post-Treatment (N = 147)
Gender		
Males	65	76
Females	53	71
Ethnicity		
White	94	117
African-American	4	1
Hispanic	13	22
Asian	5	6
Native American	2	1
Language Proficiency		
Not Proficient	4	3
Proficient	114	144
Special Education Status		
IEP	24	17
Not IEP	94	130

Implications to Internal Validity

Shadish, et al. (2002, pp. 56-61) in a general sense list eight threats to the internal validity of quasi-experiments: ambiguous temporal precedence, testing, instrumentation, selection, regression, attrition, history, and maturation. Ambiguous temporal precedence occurs when it is not obvious that the treatment preceded the effect. This threat can be

controlled for by forcing the control or pre-treatment data to come before the treatment was given.

Testing can affect internal validity when taking a test in one year influences the outcomes of the following tests. In other words, students could learn simply by taking one test and then another, which could be mistaken for a treatment effect. This could be mitigated by offering different calibrated forms of the test.

Instrumentation could be a problem if changes in the test over time lead to changes in outcomes that could mislead the researcher to believe that there is or is not a treatment effect. This may be less of a problem if the tests are calibrated every year to ensure that they are testing the same concepts and are the same difficulty each year. In addition, a test may not be sensitive to the particular treatment being studied. If so, one would expect no treatment effect, when, in fact, one may exist. Choosing a more sensitive outcome measure would diminish or eliminate the instrumentation threat, however.

Selection bias results when there are differences in groups before the treatment takes place. An example of this at the student level would be if a researcher was interested in the effects of a professional development program on math achievement and one group of students had a significantly higher proportion of English language learners, who traditionally score lower on standardized tests in math (Mazzeo, Carlson, Voelkl, & Lutkus, 2000). Student-level selection bias can be mitigated by examining student differences before moving forward with the data analysis. Teacher level selection bias could result if one group of teachers was more or less motivated to participate in the treatment. However, this could also be mitigated by allowing the teacher to serve in both

groups while student data (from two different groups of students) is used for the outcome variable.

Regression threats could occur if participants were selected because of their low or high scores on a particular measure. At the teacher level, this could occur if teachers choose to participate, because they are trying to increase their students' performance. In this case, one would then assume that their students may perform lower than other students for various reasons. However, as was the case for teacher selection bias, if the teacher is used in both the treatment and control groups, this threat could be mitigated.

Attrition occurs when "different kinds of people remain to be measured in one condition versus another," (Shadish, et al., 2002). Student attrition poses a possible threat to internal validity, because students may move in and out of schools and districts. It is then important to examine patterns of student attrition in each dataset. Teacher attrition could also be concern. For example, teachers who participated at a higher level would increase their skill-set, allowing for promotions or movement between districts. If teachers are serving as controls in both groups and they moved between measurement years, data would be lost.

History could be a factor to internal validity if external events that would lead to group differences on the outcome were to occur between measurement periods. At the teacher level, this could be something as simple as participant teachers for one treatment also participating in another treatment. The effects of the second treatment could confound the results of the first, and vice versa. This can be diminished if participant teachers are forced to only participate in one treatment at a time. In addition, curriculum changes between measurement periods could cause teachers to reform teaching practices

more or less. This is especially important when studying the link between teacher professional development and student achievement, because the effects of the program are completely reliant upon how well the participants implement the professional development concepts in their classrooms. This could be controlled for by observing classrooms at multiple times between measurement periods. At the student level, events such as violence in one year and not the other could result in spurious outcomes. To alleviate this threat, Shadish et al. (2002) recommend testing students at the same time of year. Even still, historical threats are the most difficult to control for outside a laboratory setting.

As with history, maturation is a possible threat to validity in two ways: student maturation and teacher maturation. First, student maturation could occur if students tended to gain more on the outcome measure by moving up in grade level (i.e., the gains were larger from seventh to eighth grades than they were from sixth to seventh grades). This implies that as students get older they are able to learn at a faster pace. A maturation effect could also be artificially induced by the outcome measure if the outcome ranges were different for each grade level. If so, assuming the testing measures account for differences in skill-levels from year to year, spurious outcomes could result from both the original and variation designs if teachers tended to teach lower grades during pre-treatment years and higher grades during the post-treatment years or vice versa. Normalizing the scores by grade level or including grade level as a covariate could mitigate this threat. Second, teacher maturation could occur through teacher experience, which might affect the internal validity in that teachers may simply teach better after they

have more experience. Controlling for teaching experience could diminish this threat if those data were available.

Variables

Independent Variables

One of the hypotheses for the RM-MSMSP program was that the professional development program would lead to increased student achievement. Therefore, the active independent variable for this comparison study is a two-level variable called "treatment condition". Students were divided into either the pre-treatment or post-treatment groups based on the year that they had each teacher. Pre-treatment students were in the teachers' classes the year prior to the teachers' participation in the RM-MSMSP program (in this case Spring 2005) while post-treatment students were in the classes after the teacher became treated (Spring 2007).

The RM-MSMSP program used two different reliability measures to assess how much content teachers had learned and the level at which teachers implemented what they learned. As a fidelity measure of level of implementation, the Reformed Teaching Observation Protocol (RTOP) was used on a random sample of teacher participants to evaluate the degree to which teacher instruction demonstrated reformed teaching practices. The RTOP instrument is composed of five sections, each ranging in score from 0 – 20 points, that measure lesson design and implementation, content propositional knowledge, content procedural knowledge, classroom communication, and student/teacher knowledge. The section scores were tallied for a total RTOP score that can range between 0 and 100 points, with higher scores reflecting a greater degree of reform.

The RTOP is a classroom observation protocol designed to incorporate many different standards in each of its subscales. It is comprised of 25 standards-based, inquiry-oriented, student-centered items (Sawada, et al., 2002). Results of correlations between each of the subscales and the overall score showed that the subscale scores were good predictors of the overall score (R^2 ranged from .769 to .967), and that inquiry-based reform was captured in the observational tool (Piburn & Sawada, 2000). During the pilot test of the RTOP, observers worked in pairs to get an idea of inter-rater reliability. A best-fit line through the observed ratings indicated a very high reliability (R^2 = .954). Some of these R^2 values appear to be abnormally high, so the reader is directed to pages 9 through 13 in Piburn and Sawada (2000) for confirmation and further explanation of RTOP reliability results.

To ensure inter-rater reliability of RTOP scores found in the RM-MSMSP participant teachers' classrooms, evaluators consulted the training manual (Sawada, et al., 2000) and participated as a group in the RTOP online training. The evaluators independently scored video film clips of sample teachers' lessons, compared and discussed ratings, and came to agreement on the ratings. In the field, inter-rater reliability was checked early in the project by having pairs of evaluators conduct ten initial observations. Estimates of inter-rater reliability were found to be acceptable (R^2 =.85).

Because the RTOP was not given to all participants whose data were used in this comparison study, the total RTOP score in the post-treatment year was averaged to determine the level of implementation of the professional development lessons. Sawada, et al. (2002) established basic norms for teachers by grade level and subject area. For the purposes of this study, the total norm of 48.5 for elementary and middle school and 44.1

for high school was used as a cut point between those teachers who did and did not exhibit reformed teaching practice. RTOP observations on a random sample of 34 fully and partially treated RM-MSMSP teacher participants were done over 2 years to measure their level of reformed teaching practices. Not all teachers were observed in each year, but 65% of the teachers were observed multiple times. Table 2 shows that teacher participants mostly exhibited reformed teaching practice at every school level in each year. Therefore, it appears that the teachers connected with the RM-MSMSP program do demonstrate reformed teaching practices in their classrooms over several observations. Because of these findings, it was decided not to do additional RTOP observations in 2008, but to assume that any partially or fully treated teacher would be using reformed teaching practices in his or her classroom.

Table 2

Number of RTOP observations above and below cutoff norms for each grade level during three observation periods.

	Fall 2006		Spring 2007		Fall 2007	
	n	%	n	%	n	%
Elementary[a]						
At or Above Norm	3	9	4	18	--	--
Below Norm	1	3	0	0	--	--
Middle School[b]						
At or Above Norm	24	71	15	68	7	58
Below Norm	2	6	2	9	5	42
High School[c]						
At or Above Norm	3	9	1	5	1	8
Below Norm	1	3	0	0	0	0

[a]4 elementary teachers observed in Fall 2006 and Spring 2007, but none in Fall 2007.
[b]26 middle school teachers observed in Fall 2006, 17 in Spring 2007, and 12 in Fall 2007.
[c]4 high school teachers observed in Fall 2006 and 1 in both Spring 2007 and Fall 2007.

Dependent Variables

CSAP test scores in mathematics were used to construct the student outcomes for the comparison of the original cohort control design and the variation design. The Total Scaled Score (TSS) was the dependent variable in the pre-treatment and post-treatment years for the original cohort control design. The variation on the cohort control design employed the used of the *gain* in TSS as the dependent variable.

The CSAP test is a standards-based assessment designed to measure Colorado student performance relative to the Colorado state Model Content Standards in the content areas that are assessed. CSAP tests contain both structured response and multiple-choice items. The CSAP mathematics assessment has been used in grades three through ten since 2005. Prior to 2005, the mathematics assessment was only given to students in grades five through ten.

After taking the test, students receive an overall scale score, which is designed to equate different forms of the test. This allows the score to represent the students' achievement regardless of the year in which it was administered. These scale scores are divided into four categories that represent the proficiency level of each student (see Table 3): Unsatisfactory, Partially Proficient, Proficient, and Advanced.

Extensive psychometric testing has been completed to ensure the validity and reliability of each test and the process by which it is graded. Inter-rater reliability for the constructed response questions was high for each grade level test ($\kappa > .80$). Inter-item response was also acceptable at each grade level ($\alpha > .90$). The Colorado Department of Education also engaged in extensive assessments for content and construct validity of the CSAP tests, which included reviewing the curriculum, examining each test item for

content and bias, and employing methods to minimize the variance caused by factors unrelated to the constructs measured by the test (CTB McGraw-Hill, 2007).

Table 3

Proficiency Level Scale Ranges for the CSAP Mathematics Assessment

	Proficiency Rating			
Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced
3	150 – 334	335 – 418	419 – 509	510 – 700
4	180 – 382	383 – 454	455 – 537	538 – 780
5	220 – 421	422 – 493	494 – 561	562 – 800
6	240 – 453	454 – 519	520 – 588	589 – 830
7	280 – 486	487 – 558	559 – 613	614 – 860
8	310 – 520	521 – 576	577 – 627	628 – 890

The CSAP TSS used for the original design was drawn from the pre-treatment year 2005 and post-treatment year 2007. The gain scores were created from the CSAP mathematics test TSS for the testing years 2004 through 2007. The gain scores were calculated by taking the pre-treatment or post-treatment years' scores and subtracting the previous years' scores for each student. It is important to note that, the TSS scores used in this comparison study were not normalized by grade level, even though Table 3 clearly shows that the range in scores differs by grade level. This is because the sample for this study was chosen to be one grade level only, and normalization was not necessary. However, if the sample had been made up of students in different grade levels, the TSS for each student would need to be normalized by the average score and standard deviation for his or her grade level in each testing year to account for variations in test scoring between grade levels in each test year. As Tables 4 and 5 show, if the sample of students

were made up of more than one grade level, because if a teacher teaches only sixth grade in the pretest year and only eighth grade in the posttest year, the average TSS could be larger in the posttest year simply because of the differences in the scoring range at each grade level. Tables 4 and 5 show that state-averaged gains from one grade to the next for each year can be quite different. Note that the need for normalization would only be the case for outcome variables like the CSAP test where the ranges varied by grade level or if the test were not calibrated from year to year.

Table 4

State averaged CSAP mathematics scale score at each grade level

	Grade Level				
CSAP Test Year	Fourth	Fifth	Sixth	Seventh	Eighth
2004	--[a]	509	523	539	555
2005	482	516	532	550	563
2006	489	520	529	544	562
2007	491	519	537	557	566

[a]Fourth grade not tested in Mathematics in 2004

Table 5

State averaged CSAP mathematics scale score student gains for from one year to the next year and one grade level to the next

	Grade Level Increment			
Test Year Increment	4^{th} to 5^{th}	5^{th} to 6^{th}	6^{th} to 7^{th}	7^{th} to 8^{th}
2004 to 2005	--[a]	23	27	24
2005 to 2006	38	13	12	12
2006 to 2007	30	17	28	22

[a]Fourth grade not tested in Mathematics in 2004

For the variation design, the scores were adjusted to accommodate for the possibility that the pre-treatment students may have been unusually skilled or unskilled

when compared to post-treatment classes by creating the gain scores. The main difference between the two designs is this adjustment of the outcome variable for students' prior performance through the use of the gain score in the variation design. To illustrate how the gain scores were calculated, a hypothetical example is included in the Appendix.

Covariates

There are several variables that could have a plausible effect on student achievement in addition to the professional development program. According to the National Assessment of Education Progress (NAEP) data on English Language Learners, students identified as having limited or no English proficiency score lower on reading and math assessments at all grade levels (Mazzeo, Carlson, Voelkl, & Lutkus, 2000). Similarly, the research surrounding students' with disabilities performance on achievement tests is mixed. However, there is evidence that special education status (represented by IEP status in this study) is in some way related to student achievement, whether positively or negatively (Ysseldyke, et al., 2004).

For this study, two student-level covariates, which were calculated from CSAP student demographic data, were considered prior to analysis included: English language proficiency and special education status. The English language proficiency rating was a dichotomous variable coded as "Not Proficient in English and "Proficient in English". Ultimately, less than 1% of the sample consisted of students who were not proficient in English. Therefore, it was not necessary or feasible to include it as a covariate.

Special education status was measured through the use of the individualized education program (IEP). This variable was also dichotomous and had the levels "not IEP" and "IEP". Roughly 20% of the sample consisted of IEP students and correlation

between IEP status and the dependent variable for the original cohort design was high enough to warrant some control over the variable's effects ($r = .43$). However, it was determined that there was a significant interaction between IEP status and treatment condition for both evaluation designs, breaking the assumption of homogeneity of regression slopes, and IEP status was ultimately considered an attribute independent variable.

Teacher maturation was also a concern in this study. It could occur through teacher experience, which might affect the internal validity in that teachers may simply teach better after they have more experience. This could be even more evident in the context of the middle school mathematics professional development program, especially in Colorado, as teachers are not required to be certified in middle level math. Although the literature suggests a complex relationship between student achievement and teacher experience, maturation seems to at least have some effect on student achievement, and should be considered as a threat to validity. The teaching experience variable that was initially considered was based on teacher self-reported years of experience during their pre-treatment index year. It was coded into two levels based on findings from the literature that most student achievement effects come from the early part of a teacher's career (Boyd, et al., 2007; Darling-Hammond, 1999). Thus the levels of teacher experience were coded as follows: "1-5 years" and "6 or more years". Ultimately, the correlations between teaching experience and the dependent variables of both evaluation designs were not significant, and teaching experience was not included in the final analyses.

Data Collection and Management

The CSAP data collection protocol was approved by the IRBs at both UCD (the administrator of the RM-MSMSP grant) and Colorado State University (the evaluation subcontractor for the grant). Data collection began in the summer of 2006, using the following procedure. Three types of data were collected from the participating school districts: teacher data, student data, and course level data. The teacher data included RM-MSMSP participation information (i.e., courses taken and years participated) and self-reported survey data (i.e., years of teaching experience). The survey data came from a project-administered demographics survey that teachers voluntarily completed every year. Student data were transmitted to the RM-MSMSP project coordinator during the fall of each year from the school districts in the form of CSAP data through Microsoft Excel. Other than a unique student ID given by each school district, the data were cleared of student identifiers by the project coordinator before transmission to the evaluators. After this point, only the school district was able match the student ID with the actual student, creating student anonymity on the part of the evaluators.

The data stored in these files included overall mathematics TSS, proficiency level, and demographic information (i.e., language proficiency, free-reduced lunch status, and IEP status). The course level data were provided by the school districts in the form of course lists, called crosswalk files. Each district was given a list of teachers who had participated in the RM-MSMSP program and the years in which they participated. The districts then matched the students in each teacher's courses to the teacher. They stripped all student identifiers away except for the unique numerical ID. Then the district liaison sent the crosswalk files to the RM-MSMSP evaluator. The crosswalk files include the

teacher name, the student IDs for each teacher's classes, the grade level of each course, and the subject taught (i.e., math, science, or elementary). The latter variable allowed for the inclusion of only those teachers who actually taught a math course in the year of interest.

These district by grade level data were stored in Microsoft Excel files in a password protected folder on a secure server. The crosswalk files from each district and an additional file containing teacher-level data were also stored as Excel files in the same folder. Because of the large number of Excel files, a Microsoft Access 2007 database was used to query the data, creating a final aggregated dataset for each year. Before data were input into the database, a two-step manual aggregation process prepared the individual Excel files for use in queries. Although these two steps were not entirely necessary for the database aggregation process, they aided in diagnosing problems. One such problem was that the files did not always come in the same format and had to be reformatted.

First, district-level CSAP and crosswalk files for each year from each of the grade-level were manually created by pasting the data from each of the original files into a yearly district-level file. For example, if there were two years of data and six districts, there would be two Excel files for each district in each year for each district, resulting in 12 total files each year. Next, the district-level files were pasted into a master CSAP file and teacher file for each year (i.e., for one year of data there would be two Excel yearly files, one CSAP and one crosswalk file). Then several queries were run in Microsoft Access 2007 on the yearly Excel files that resulted in an aggregated, ready-to-analyze dataset for each year. The aggregation procedure was as follows:

1. Identify teachers who were sixth grade elementary math teachers and have been fully treated in math.
2. Using the yearly crosswalk files, pull the student identifiers for these teachers' pre-treatment index year and post-treatment index year from the related yearly crosswalk files.
3. Using the yearly math CSAP files, get the CSAP scores and demographic information for each student in the pre-treatment index year (Year 1).
4. Find these same students in the math CSAP yearly file for the year that precedes the pretest index year (Year 2), and pull their math CSAP data. Note: Only keep scores from students who appear in both years' files.
5. Subtract the Year 2 score from the Year 1 score to get a gain score for each student.
6. Repeat these steps for the post-treatment index year.

Note that the previous procedure was performed for the variation design. However, all the above steps except for (4) and (5) were done to create the dataset for the original cohort control design analysis.

Data Analysis

After employing the aggregation scheme described above, two datasets were produced for each design, one pre-treatment dataset and one post-treatment dataset for each year, with the outcome variable dependent on the evaluation design (i.e., TSS score for the original cohort control design and gain score for the variation design). These datasets were combined and transferred manually to a separate file in SPSS (Version 17.0) to be analyzed. Each student had one row in the SPSS data table and was classified

as either pre-treatment (0) or post-treatment (1), which is consistent with a between-groups design.

Before beginning the inferential analysis for each year, descriptive statistics were run on both the pre-treatment and post-treatment conditions to assess whether or not there were large differences in the student groups. As was presented earlier, the pre-treatment and post-treatment groups did not significantly differ in their demographic make-up. In order to give a real-world means of comparison for the two research designs, a 2 x 2 Analysis of Variance (ANOVA) was run with treatment condition and IEP status as the independent variables. The dependent variables were CSAP TSS and CSAP gain in TSS for the original and variation designs, respectively. The interest in running the 2 x 2 ANOVA was to investigate whether the real data outcomes –the effects of the professional development program and especially the suspected interaction between IEP status and treatment condition—supported the theoretical advantages of the variation design. In other words, this analysis was done to make sure that the analysis outcomes made conceptual sense.

CHAPTER 4 – DISCUSSION, IMPLICATIONS, AND CONCLUSIONS

The objective of using the variation of the cohort control design in the RM-MSMSP evaluation was to add a control for students' previous performance, thus improving the causal statements made about the program effects on students' mathematics achievement. This chapter attempts to answer the four sub-questions posed in the first chapter, and in doing so, determine if the variation design offers another solution for evaluating professional development effects on student achievement.

This chapter is divided into three main sections. First the theoretical and real-data comparisons between the two evaluation designs are presented in the discussion section. This section is meant to answer the first two questions posed in Chapter 1. Then the two designs are compared from a practical, utilitarian standpoint in order to answer the second two questions posed in the first chapter. Finally, Chapter 4 is concluded by offering suggestions for further research, answering the main question of whether or not the variation design is a new methodological solution to the problem of practice, and summarizing the advantages and disadvantages of the variation design for evaluating professional development effects on student achievement.

Discussion of Design Comparisons

As was noted in Chapter 3, Shadish, et al. (2002) give three means of better controlling for alternative plausible explanations to outcomes obtained through quasi-experimentation. When comparing the two evaluation designs presented in this study, it is important to explore how the two evaluation designs compare on these three principles

for quasi-experimental design control and how well the designs control for other plausible explanations. This section is broken down into design comparisons based on the three principles for ruling out alternative explanations.

Internal Validity Threats

Ambiguous temporal precedence. In this study, both designs incorporated data on students before teachers began participation in the RM-MSMSP program as the pre-treatment or control group, while specifically choosing post-treatment students who were in teachers' classes after the teachers became treated. By doing this, the treatment was forced to precede the effect. Just as for the case of the RM-MSMSP data, inherent in both the original and variation designs is the idea that the control always precedes the treatment which is always given before the effect, if one exists. Therefore, this particular threat to internal validity is neither an advantage nor disadvantage to either design.

Testing. The threat of biased effects obtained through students learning through test-taking as opposed to learning through the given treatment is especially salient when using a design that incorporates a pretest and a posttest, such as the variation design. Because the original design solely relies on a posttest for each cohort, testing is not a threat, which is an advantage of the original design.

For the RM-MSMSP program data, the outcome measure that was used was a calibrated test that was given in each grade level. The difficulty of the content being tested was different for each grade level as new material was learned, while the overarching concepts were the same. In other words, students take a completely different test in fifth grade from the one they take in sixth grade. Also, 10% of the test questions for each form of the test were replaced with new questions and recalibrated. In this way,

testing was not seen to be a threat to the variation design in this study. However, this is a valid threat for other applications of the variation design, and care must be taken to rule out the possibility that outcomes were not due to learning through testing.

Instrumentation. Instrumentation can bias results if the forms of the test are any more or less difficult from one year to the next. This is a salient threat to both designs, because changes in test difficulty could affect the comparison of a single score or a gain score between two groups. For example, if students in the pre-treatment group were to have a more difficult test form than students in the post-treatment group, then there is an increased threat of making a type I error. Similarly, if the test changes every year, then students in the same pre-treatment cohort might have much smaller gains than students in the post-treatment cohort, which would have the same biased effect. One way to mitigate this threat is simply to use the same test for the pre-treatment and post-treatment groups. If this is done with the variation design, the testing threats discussed previously might then become an issue. The other way to reduce this threat without using the same test is to calibrate the outcome measure to ensure that it is testing the same concepts and is the same difficulty. It might seem at first like the original design has an advantage over the variation design here, because it more easily allows the use of only one test. However, in studies of the effects professional development on student achievement, the outcome measure is often a state test, such as the one used for the RM-MSMSP program, or other test that is different each year, but is calibrated. Therefore, this threat is neither an advantage nor disadvantage to either design if attention is paid to the outcome measure of choice.

Another instrumentation threat comes from the sensitivity of the measure to the treatment. For example, if one is examining the effects of an inquiry-based program on student achievement, but chooses an outcome measure that is based on memorization only, the measure will most likely not measure an effect. This increases the chances of making a type II error. This threat is salient for both designs, as an insensitive measure is likely to not show differences in scores or gain scores. In fact, for the RM-MSMSP data, there was some concern that the CSAP test would not be sensitive to the inquiry-based professional development. However, the professional development was also highly focused on increasing teacher content, which if translated to the classroom, might show an effect on the test even if there was a concern over the lack of inquiry-based skills being tested. This example makes it clear that, despite the evaluation design, attention must be paid to the outcome measure before choosing one in a study.

Regression. One design modification made to both the original design as depicted in Shadish, et al. (2002) and the variation design used two types of data, teacher and student data. In doing so, teachers were forced to act in both the pre-treatment and post-treatment groups, acting as teacher-level controls. Therefore, for this study, teacher regression effects were not considered to be problematic. This was not a unique feature of either design, and is not considered an advantage or disadvantage to using either design. However, it is important to note that if the teachers in the pre-treatment and post-treatment group were not forced to be the same, then regression effects should be investigated by examining the teacher participants.

Attrition. Teacher attrition is a concern for both the original and variation designs if the teacher is being used as a control in both the pre-treatment and post-treatment

groups, because they need to continue teaching in a location where data could still be collected on their students in order for their class data to be included for use in a study. In this comparison study, many teachers who became treated early on in the program moved out of their original district or were promoted to an administrative (non-teaching) position. This was a problem, because the districts were so different in terms of demographics and policies, and using pre-treatment data from one district and post-treatment data on another would add another level of uncertainty to the outcomes. It would seem that these phenomena could be quite common, because teachers who have a great deal of participation in professional development would increase their skill sets, allowing them to compete for higher-salaried teaching and administrative positions. If this occurs, data on these teachers' students would not be included in the analysis using either design, and the outcomes could be underestimating the effects of the program. Obviously, this would depend on the data being collected and the manner in which it is collected, but teacher attrition threats are salient to any study and could be problematic for either the original or variation design.

Student attrition is troublesome for the variation design, because two years of data for each student is necessary to create the gain scores used in the design. However, if the student stays within one district (or within an area where data are being collected), the variation design can accommodate for their movement. To do this, gain scores can only be calculated on students who appear in both the pre-treatment (or post-treatment) year and the previous year were used. In other words, all students must have a pretest and posttest measure in order for them to be counted in the overall analysis. Even if this is done, student attrition could be a problem for the variation design, because students who

do not have both pretest and posttest measures are not included in the analysis. This might occur if students moved into or out of the data collection area during the pretest or posttest year. Student attrition is especially a threat to the variation design if students of different abilities have different rates of attrition in the pre-treatment and post-treatment groups. This is an advantage to using the original design, because students only need to be present in their respective year (pre-treatment or post-treatment). However, *if data are available on all students*, it is fairly simple to ensure that the gain scores are only calculated on students who have both a pretest and posttest, making the variation design as usable as the original design. This is usually not the case, however.

Maturation. There are two types of maturation threats to consider when doing studies on the effects of professional development on student achievement: student and teacher maturation. Student maturation is a threat to both designs. If students score higher on the outcome measure simply because they are older, then this could bias the results of the original design. If this phenomenon were true, then the gains from fifth to sixth grade might be different than those from sixth to seventh grade, which makes the maturation threat possible for the variation design. This implies that as students get older they are able to learn at a faster pace. One way to mitigate this problem is to create a sample of one grade level, as this study does. It was appropriate to do this for this study, because it was only a comparison of two evaluation designs. But, in an actual evaluation, using data from only one grade level may not be practical or useful. Another way to weaken this threat is to use an outcome measure that is scaled in such a way as to account for the assumption of students score higher each year as a product of age. It should be noted,

however, that student maturation is only a problem if students in the pre-treatment and post-treatment group differ significantly on their age or grade-level make-up.

Teacher maturation could also be problematic in that the treatment effects could be confounded by the fact that teachers may get better at teaching over time. This is a problem for both designs, because the pre-treatment and post-treatment years are defined by when the treatment is completed in both designs. Although the variation design requires four years of data, the pre-treatment and post-treatment years are the same for the original and variation designs. This means that teacher participants would have the same amount of time to mature between the pre-treatment and post-treatment years in both designs. Therefore, students in the post-treatment group could score higher or have larger gains than those in the pre-treatment group simply because the teacher had longer to gain experience before teaching the post-treatment students. Therefore, teacher maturation is a threat that should be carefully considered when using either the original or variation evaluation design. As was mentioned previously, teacher experience data were analyzed before the main analysis to see whether or not teacher experience was correlated with the outcome measures. In this study, they were not significantly correlated, and were not controlled for. However, because this threat pertains to both designs equally, it is important to investigate how teaching experience is related to outcomes in each study.

History. As with maturation, history threats are equally valid for both the original and variation designs. Because both designs use data collected from at least two points in time, there is the threat that events took place at one point in time, but not the other. Although the threat is valid for both designs, it is much more of a concern variation design, because it employs four years of data. This is especially true in terms of

professional development programs, as it is often difficult to determine if teachers are participating in more than one program. If so, how can the any effect confidently be attributed to any single program? This was a valid threat for the RM-MSMSP data used in this study, as participants were not discouraged from participating in other programs. Similarly, events that might affect students could be different from year to year, confounding student effects.

Shadish, et al. (2002) contend that there are two ways to reduce the historical threat to validity. The first is to select groups from the same general area. This is fairly vague, as a general area could be a school, district, or town. The Denver-Metro area was taken to be the geographical area from which the analysis was done. Second, they suggest that testing be given at the same time for all participants. While students in the RM-MSMSP sample were tested at the same time of year due to state testing practice, the pre-treatment group still took the state tests at least one year before the post-treatment group. This leaves some uncertainty as to historical events that may take place in between. The basic assumption was that historical events occurred randomly in time for each participant/school. Therefore, if this assumption was not true, history could be a very large threat to validity, because other measures were not in place to attempt to account for possible events (e.g., records of school violence or other professional development activities).

Selection bias. Much like maturation and history threats, selection bias threats involve both teachers and students. Teacher selection bias occurs when teachers in one group are selected for their ability level. This threat is easily accommodated in both the

original and variation designs by allowing teachers to serve as their own controls. In other words, the same teachers are in the pre-treatment and post-treatment groups.

Student selection bias is more difficult to deal with. The only way to lessen selection bias in the original design is to examine the demographics to make sure the students are similar in each group. However, there may be student differences that cannot be accounted for through demographics alone. In fact, the inclusion of a pretest measure to allow for the inclusion of students' prior achievement is specifically identified by Shadish, et al. (2002, p. 151) as an improvement to the original design. The variation design adds this control in the form of a gain score.

Design-Implemented Control

Shadish, et al. (2002) explain, "By adding design elements...quasi-experimentation aims either to prevent the confounding of a threat to validity with treatment effects or to provide evidence about the plausibility of those threats," (p. 105). The variation design does just that by adding a pretest measure in both the pre-treatment and post-treatment groups. In fact, Shadish, et al. (2002, p. 151) give an example a of mixed (between/within) design. The variation design compared in this study takes the mixed design one step further by creating a gain score between the pretest and posttest, making the overall design a between-groups design with an element of a within-groups design.

While it was previously discussed in the theoretical sense that the variation design would have certain advantages over the original design (i.e., the control for prior student performance), it was not certain if the analyses would produce meaningful outcomes. In other words, a design that makes sense theoretically but does not work in reality is of no

practical value for an evaluation. Therefore, two separate analyses were done, one for each evaluation design, so that the outcomes could be compared for their meaningfulness. Tables 6 and 7 show the results of the 2 x 2 ANOVAs that were run for each design.

Table 6

Adjusted mean CSAP math scores (original design) and gain scores (variation design), standard deviations, and sample sizes for the main effects groups

		Original[a]		Variation[b]	
	n	*M*	SD	*M*	SD
Treatment Condition					
Pre-Treatment	117	520.98	86.86	28.69	46.40
Post-Treatment	147	505.00	108.03	13.51	57.59
Special Education Status					
IEP	224	465.45	70.04	20.50	37.27
Not IEP	40	560.53	69.89	21.70	37.25

[a]Dependent variable is student CSAP math score.
[b]Dependent variable is student CSAP math gain score.

The RM-MSMSP data were hypothesized to show a difference in mean scores between the pre-treatment and post-treatment groups when using the original cohort control design. It was expected that if the program met expectations, the students in the post-treatment group would have higher scores than those in the pre-treatment group. Similarly, it was hypothesized that students in the post-treatment group would gain more than those in the pre-treatment group when using the variation design. IEP status was included, because there were obvious differences in how the professional development affected IEP students in the pre-treatment and post-treatment time groups. Overall, IEP was thought to lower student scores in general, but there was no directional hypothesis for gain score. Additionally, it was assumed that if the outcome measure was not

sensitive to the treatment, there would be no effect for either the original or variation design outcomes.

Table 7

Comparison of Analysis of Variance for sixth grade CSAP math achievement using the original cohort control design and the variation design.

		Original[a]		Variation[b]	
	df	*F*	partial η	*F*	partial η
Overall Model	3	22.60***	.21	5.67***	.06
Main Effect					
Treatment Condition	1	1.78	.01	5.63*	.02
IEP Status	1	62.84***	.20	.04	.00
Interactions					
Treatment Condition X IEP Status	1	6.25*	.02	16.34***	.06
Error	263				

[a]Dependent variable is student normalized CSAP math score (R^2 = .21).
[b]Dependent variable is student normalized CSAP math gain score (R^2 = .06).
*$p < .05$; ***$p < .001$

The hypothesis that post-treatment students would score higher than pre-treatment students was not supported. After looking at the data, it was found that this was not due to the original cohort control design features or to a lack of sensitivity of the measure, but most likely due to a large interaction between treatment condition and IEP status. In fact, Figure 2 shows that non-IEP post-treatment students scored higher, on average, than those in the pre-treatment year, while IEP post-treatment student scored lower than those in the pre-treatment group. However, even after separating students by IEP status, non-IEP students in post-treatment group did not significantly differ (t = -1.55, p = .12) on their math scores from the pre-treatment non-IEP students. There was also not a significant treatment (t = 1.73, p = .09) effect for IEP students. The interaction occurred

because the spread in scores between IEP and non-IEP students were larger in the post-treatment group than in the pre-treatment group.

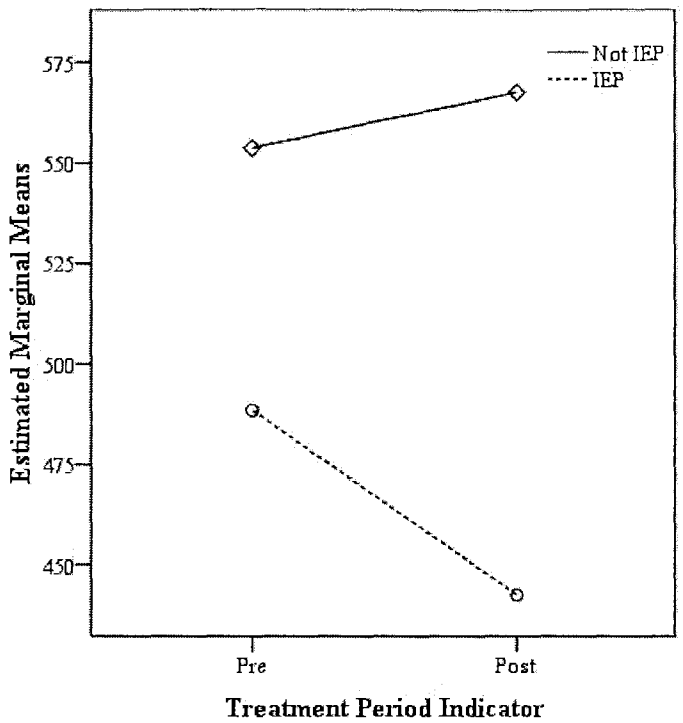


Figure 2: Adjusted means for sixth grade student math scores (original design) during the pre and post-treatment years. Lines represent IEP status.

Do the previous results indicate that the treatment did not have an effect on either group individually or were there possible differences in students that were not accounted for with the original design? To help answer this question, the variation design added one more year of data to each group to help account for their prior achievement. Tables 6 and 7 show that there was a significant treatment effect on gain scores where post-treatment students have smaller gains, on average, than the pre-treatment students. However, there was a significant interaction between IEP status and treatment condition. Figure 3 shows the interaction effects for the variation design.

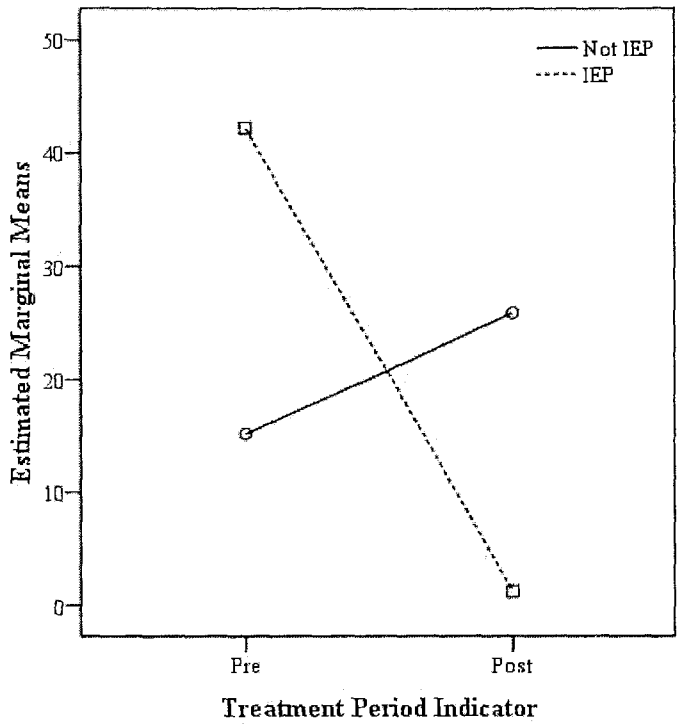


Figure 3: Adjusted means for sixth grade student math gain scores (variation design) during the pre and post-treatment years. Lines represent IEP status.

The first obvious pattern in Figure 3 is that, just like for the original design, the treatment appears to have the opposite effects on IEP students than it has on non-IEP students. In fact, when separated into groups by IEP status, non-IEP students in the post-treatment group (M = 25.85, SD = 36.50) had significantly (t = -2.16, p = .03) larger gains than those in the pre-treatment group (M = 15.16, SD = 36.58). This can be interpreted as indicating that, after controlling for students' prior performance, students in the post-treatment group gain almost 10 more points from year to year than pre-treatment students. The same cannot be said for IEP students, and this was the reason for the negative treatment effect. IEP students in the post-treatment group (M = 1.18, SD = 34.02) gained significantly (t = 3.32, p = .002) less than those in the pre-treatment group (M = 42.22, SD = 41.68).

From a professional development perspective, the effects on IEP students may not be ideal. However, from a design perspective, these outcomes are still realistic. Most likely, the large gains in the pre-treatment IEP students were due to regression to the mean effects, because they had much smaller scores in the pretest year than the non-IEP students, and had a larger range from which to score in the posttest year. The same might have been true for the post-treatment group, but there could have also been overarching professional development effects. It is reasonable to assume that as teachers gain more math content knowledge and pedagogical content knowledge, they will begin to teach more difficult material in a manner that may not be conducive to students needing special attention. Evidence of reformed teaching practice was presented in Chapter 3, so this explanation is believable.

The interactions in the variation design indicate that the CSAP test was sensitive to at least part of the professional development, because there were significant treatment effects found. The same effects were not found in the original design, which points to the design control of the variation design adding an advantage for its use in professional development evaluations. Further, the fact that the variation design gives reasonable outcomes confirms, from a practical standpoint, that the design is usable.

Complex Pattern Matching

Shadish et al. (2002) explain that complex pattern matching is when "a complex prediction is made about a given causal hypothesis that few alternative explanations can match," (p.105). This principle is not unique to either evaluation design compared here, but it does apply to both of them. From a research perspective, complex pattern matching might mean that the researcher makes hypotheses about suspected interactions between

variables. Although complex hypotheses are meant decrease the number of alternative explanations, there is no reason to make them if they do not (Shadish, et al., 2002). Therefore, for both the original and variation designs, it is important to make numerous specific and varied hypotheses if they will help rule out alternative explanations.

Practical Implications for Choosing Between Designs

While working with both evaluation designs, it became evident that although the variation design added further controls to the original design, it was more difficult to employ given the need for at least four years worth of data. One of the main problems with this is that it was often difficult to get teacher and student data for all four of the years necessary. Most of the data were historical, and not all the data were in the same format. Therefore, the partner districts had to be extremely cooperative in reformatting and sending the data. Further, the use of four years of data created complexities concerning the use of the database that became necessary for employing the variation design. Instead of simply finding each teacher's pre-treatment students and their scores and doing the same for one year of post-treatment scores, this new database had to go one step further by also finding the same students in the years prior to their pre-treatment and post-treatment years, respectively, making sure that only students who had scores in both years were included. In addition, the creation of the gain score added another step to the data aggregation and formatting process.

The addition of the gain score also added some concerns about the reliability of the new outcome measure. As was shown in Chapter 2, the use of gain scores in general can be problematic, because of possible bias and unreliability of the measure itself. Zimmerman and Williams (1998) discussed gain score reliability, and found that as long

as the posttest variance had not decreased, the gain score could be considered a reliable measure. In this comparison study, the variance of the posttest was larger than that of the pretest in the post-treatment group. However, the posttest variance was smaller than that of the pretest in the pre-treatment. This would indicate the possibility that the outcomes are not reliable. In cases such as this, the variation design would not be appropriate to use, and the original design would be a better.

In addition, bias and unreliability would be a concern if the gain scores were calculated based on two different groups of students, which is often done. For example, if the pretest measure was from 6th grade students in 2004 and the posttest measure was from a different set of 6th grade students in 2005, the gain score would not be a true measure of student gain from one year to the next. In addition, if pretest and posttest measures came from the same group of students at two points in time, but no care was taken to make sure that all the students from the pretest group were also in the posttest group, bias could occur. However, these concerns are not germane here, because all students appeared in both years used to calculate each gain score by connecting each student's score in one year to his/her score in the previous and not including students who had a score in only one of those years.

Hake (1998) suggests that a ceiling effect could occur if most of the students had high scores on the pretest measure (scores for the years prior to the pre-treatment and post-treatment years, respectively). If this were the case, there might be no effect found simply, because students had a smaller range on which they could improve. To be sure there were no ceiling effects in this comparison study, the student outcomes for the years prior to the pre or post-treatment years, were plotted to confirm that neither group's

outcomes were high compared to the range for the CSAP test. Figure 4 shows the average student TSS at all grade levels for the pre-treatment and post-treatment groups in the year directly prior to the pre and post-treatment years, respectively. These scores are bounded by the minimum and maximum CSAP scores for each grade level. Neither the pre-treatment nor post-treatment groups had significantly higher or lower pretest scores. In addition, the average pretest scores for both groups fall roughly halfway between the upper and lower bounds of the TSS for each grade level.

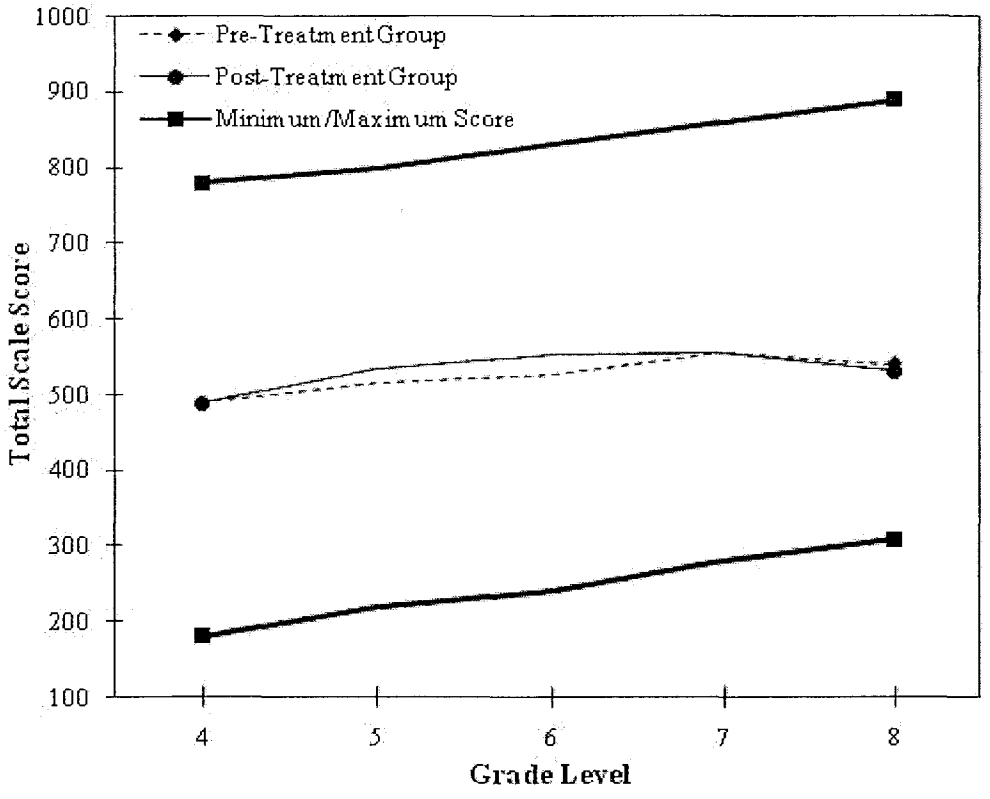


Figure 4: Student TSS in the year directly prior to the pre-treatment and post-treatment year, at each grade level. Bolded lines represent the minimum and maximum TSS at each grade level.

Because this was a comparison study, the concerns about unreliability were kept, although the analyses were still done for the comparisons. The other bias and unreliability

concerns were not considered to be supported for this comparison study. However, this finding is dependent on the study, and would need to be tested for each individual evaluation. Therefore, it is important to examine the possibility of a ceiling effect, and if one exists, the original design would be preferable to the variation design. This illustrates that while the introduction of a gain score adds controls to the internal validity of the outcomes, it also adds complications to the use of the design. In terms of gain scores, it becomes a trade-off between controls and outcome measure reliability.

From a research perspective, the variation design presented and discussed in this dissertation brings with it several expectations about the project being evaluated, including the following: (1) there is access to at least four years of historical data, (2) there are a large number of participants who teach students who will ultimately make up the sample, and (3) the teachers do not move into or out of districts between pre-treatment and post-treatment years. Realistically, there tends to be a large amount of attrition associated with programs like RM-MSMSP, both into and out of the program and also within and between districts. Therefore, although there may be large initial numbers of participants, the actual final count could be significantly lower. Attrition is problematic for any design, but because the variation design ties teachers to students and creates student gain scores, it is particularly important to pay attention to attrition. Fewer teacher participants would lead to a reduction in the sample size and could lead to an unrepresentative sample.

Recommendations for Further Research

This study explored the feasibility of employing a variation on an already existing evaluation design for use specifically in large-scale, complex professional development

programs. Comparisons showed that there were advantages and disadvantages to using either design. If the requirements in the data are met, it would be interesting to employ the variation design in an actual program evaluation. One suggestion would be to use the data already obtained through the RM-MSMSP program. Although there was some concern about the sensitivity of the CSAP test to inquiry-based professional development, there are other possibilities about the interpretation of the results that should be discussed as it would add to the small body of literature associated with professional development effects on student achievement. In addition, repeating the evaluation using the independent variables suggested here and other covariates found in the literature along with a different measure, preferably designed to measure inquiry-based learning, would shed light on the amount of variance being accounted for by the model for the variation design. If, in fact, the measure is reliable and the amount of variance being accounted for is still low, it would point to extraneous variables that also need to be explored, and it would confirm observations brought up by (Wayne, et al., 2008) about the difficulty of evaluating professional development related to student achievement.

Conclusions

This study focused on the comparison of two evaluation designs, the original cohort control design and a variation on that design. The first two research sub-questions explored the methodological implications of using the variation design compared to the original cohort control design. Comparisons of the threats to internal validity for both designs showed that the introduction of the pretest through the use of the gain score adds controls for possible selection bias that would still be a threat for the original design. Thus, the gain scores are an advantage for the use of the variation design if certain

conditions of the data are met: (1) the variance of the posttest measure is not smaller than the pretest variance, (2) gain scores are calculated only for students who have both a pretest and posttest measure, and (3) no ceiling effect exists in the data. If one or more of these conditions are not met, then concerns about the unreliability of the gain score measure itself should be considered and addressed. These conditions bring to light certain advantages of the original cohort control design, in that there is no pretest measure. This makes the design simpler, and threats such as testing are not a problem. However, there are still methodological disadvantages to the original design, namely selection bias.

The comparison of the two designs showed that, as expected, the outcomes were different, but both sets of outcomes made real-world sense. In other words, the theoretical discussion about the advantages and disadvantages to each were supported, because the data were able to be explained.

The second two sub-questions were related to the utility of the variation design for the evaluation of similar professional development programs. The main conditions that need to be met when using the variation design are (1) the access to at least four years of student *and* teacher data, (2) the ability to create and manage a large dataset to query and aggregate the data, and (3) data that do not lead to a ceiling or floor effect, which would negate the value of using gain scores. If these conditions are met, the variation design is no less practical to employ than the original cohort control design. In fact, because both the original cohort design and the variation design rely on the time it takes for teachers to go from not treated to treated, the variation design is more practical as it adds to the internal validity of the findings. Access to a database program that allows

for data querying is also beneficial for both designs, but is not necessary. It simply reduces the amount of time it takes to aggregate the data.

In addition to the methodological and practical advantages of the variation design, it has advantages from a policy standpoint. The variation design uses student gains, or a change in score from one year to the next, as the outcome measure, which is similar to the value-added research designs that are becoming more popular amongst district administrators in determining the quality of school performance. In this way, the results of any evaluation that uses the evaluation design may be more pertinent when presenting to program and district administrators, as it the outcomes are more concerned with student change over a short period of time instead of simply differences in scores between two groups of students.

Given the highly charged political climate surrounding student achievement and teacher professional development, there is much to be gained from research that can rigorously address the relationship between high quality professional development and student learning. The existing literature (e.g., Yoon, et al., 2007) shows that professional development designed with high quality components (i.e., inquiry-focus, content-based, continuous, etc.) can lead to an increase in student achievement. However, this consensus of how professional development should be designed is fairly informal in the literature, with little evidence on the exact effect on student achievement. There are numerous findings on the lack of the depth in current research designs for studying the connection between professional development and student achievement (Muijs & Lindsay, 2008; Wayne, et al., 2008). The variation design compared here was specifically proposed to address the need for more depth in the program evaluation of the RM-MSMSP program

and could be used in other program evaluations to help fill the gap in the literature concerning the link between professional development and student achievement. Realistically, however, the choice of evaluation design is one of trade-offs. Given the above methodological and practical conditions are met and the outcome measure is calibrated, the variation design would be a good choice for a professional development evaluation.

REFERENCES

Ash, D. (2000). The process skills of inquiry. In *Foundations: Inquiry - Thoughts, views, and strategies for the K-5 classroom* (pp. 51-62). Washington, D.C.: National Science Foundation.

Ball, D. L., & Cohen, D. K. (1999). Developing practices, developing practitioners: Toward a practice-based theory of professional development. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the Learning Profession: Handbook of policy and practice* (pp. 30-32). San Francisco: Jossey-Bass.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433-449). New York, NY: Macmillan.

Basista, B., & Matthews, S. (2002). Integrated science and mathematics professional development programs. *School Science and Mathematics, 102*(7), 359-370.

Birman, B., LeFloch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., et al. (2007). *State and local implementation of the No Child Left Behind Act: Volume II - Teacher quality under NCLB Interim report.* Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Programs Studies Service.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 30*(8), 3-15.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*, 417-436.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wykoff, J. (2007). *The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools*: CALDER Working Paper.

The Brookings Institution. (1998). *Learning what works: Evaluating complex social interventions*. Washington, D.C.

Carpenter, T.P., Fennema, E., Peterson, P.L., Chiang, C.P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499-531.

Charmaz, K. (2006). *Constructing grounded theory*. London: Sage Publications, Inc.

Chen, H.-T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage Publications, Inc.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice, 20*(4), 19-28.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

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

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

CTB McGraw-Hill. (2007). *Colorado student assessment program: Technical report 2007*. Monterey, CA: The McGraw-Hill Companies, Inc.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of the state policy evidence*. Seattle: University of Washington, Center for the Study of Teaching and Policy. Document No. R-99-1.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57-77.

Desimone, L., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educationa Evaluation and Policy Analysis, 24*(2), 81-112.

Dyasi, H. (2000). What children gain by learning through inquiry. In *Foundations: Inquiry - Thoughts, views, and strategies for the K-5 classroom* (pp. 9-14). Washington, D.C.: National Science Foundation.

Fetler, M. (2001). Student mathematics achievement test scores, dropout rates, and teacher characteristics. *Teacher Education Quarterly, 29*(1), 151-168.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90*(3), 560-569.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 122-145.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.

Guskey, T. R. (2003a). Results-oriented professional development: In search of an optimal mix of effective practices. In A. C. Ornstein, L. S. Bhear-Horenstein & E. F. Pajak (Eds.), *Contemporary issues in curriculum* (pp. 321-333). Boston: Allyn and Bacon.

Guskey, T. R. (2003b). What makes professional development effective. *Phi Delta Kappan, 84*(1), 748-750.

Guskey, T. R., & Sparks, D. (2002). *Linking professional development to improvements in student learning*. Paper presented at the Annual Meeting of the American Educational Research Association.

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*(1), 1-29.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1*(1), 19-41.

Hanushek, E. A. (1996). A more complete picture of school resource policies. *Review of Educational Research, 66*(3), 397-409.

Hanushek, E. A. (2006). Teacher Quality. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 3-29). New York, NY: North Holland Publishers.

Hawley, W. D., & Valli, L. (2001). The essentials of effective professional development: A new consensus. In D. Boesel (Ed.), *Continuing professional development* (pp. 1-17). Washington, D.C.: U.S. Department of Education, National Library of Education.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind Policy Initiative. *Educational Evaluation and Policy Analysis, 29*(2), 95-114.

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

Huffman, D., Thomas, K., & Lawrenz, F. (2003). Relationship between professional development, teachers' instructional, and the achievement of students in science and mathematics. *School Science and Mathematics, 103*(8), 378-387.

Huziak-Clark, T., Van Hook, S. J., Nurnberger-Haag, J., & Ballone-Duran, L. (2007). Using inquiry to improve pedagogy through K-12/University partnerships. *School Science and Mathematics, 107*(8), 311-324.

Joyce, B., Calhoun, E., & Hopkins, D. (1998). *Models of teaching: Tools for learning.* Buckingham: Open University Press.

Kennedy, M. M. (1998). *Form and substance in inservice teacher education.* (Research Monograph No. 13). Madison, WI: National Center for Improving Science Education.

Kimmel, H., Deek, F. P., Farrell, M. L., & O'Shea, M. (1999). Meeting the needs of diverse student populations: Comprehensive professional development in science, math, and technology for teachers of students with disabilities. *School Science and Mathematics, 99*(5), 241-249.

Laczko-Kerr, I., & Berliner, D. C. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10*(37), 1-69.

Lasley, T. J., Siedentop, D., & Yinger, R. (2006). A systematic approach to enhancing teacher quality:The Ohio model. *Journal of Teacher Education, 57*(1), 13-21.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educationa Evaluation and Policy Analysis, 15*(2), 129-152.

Loucks-Horsley, S., Love, N., Stiles, K. E., Mundry, S., & Hewson, P. W. (2003). *Designing professional development for teachers of science and mathematics* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Loucks-Horsley, S., & Matsumoto, C. (1999). Research on professional development for teachers of mathematics and science: The state of the scene. *School Science and Mathematics, 99*(5), 58-72.

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. *Sociological Methodology, 18*, 347-409.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (No. 2000-473). Washington, DC: Department of Education, National Center for Educational Statistics.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standard-based high school mathematics. *Journal for Research in Mathematics Education, 32*(5), 493-517.

McDiarmid, G. W., David, J. L., Kannapel, P. J., Corcoran, T., & Coe, P. (1997). *Professional development under KERA: Meeting the challenge.* Lexington, KY: The Partnership of Kentucky Schools & The Prichard Committee for Academic Excellence.

Meyer, S. J., & Sutton, J. T. (2008). *Examining teacher outcomes and student mathematics achievement outcomes in the Math in the Middle (M^2) Institute Partnership.* Paper presented at the Annual Meeting of the American Educational Research Association.

Mohr, M. J. (2006). An assessment of middle grades preservice teachers' mathematics knowledge for teaching (Vol. 68).

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

Monk, D. H., & King, J. A. (1994). Multilevel teacher resource effects in pupil performance in secondary mathematics and science: The case of teacher subject matter preparation. In R. G. Ehrenberg (Ed.), *Choices and consequences: Contemporary policy issues in education* (pp. 29-58). Ithaca, NY: ILR Press.

Muijs, D., & Lindsay, G. (2008). Where are we at? An empirical study of levels and methods of evaluating continuing professional development. *British Educational Reserach Journal, 34*(2), 195-211.

Muijs, D., & Reynolds, D. (2000). *Effective teaching: Evidence and practice*. Londong: Paul Chapman Publishing.

National Center for Education Statistics. (2006). *The nation's report card: Mathematics 2005*. Washington, D.C.: U.S. Government Printing Office.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, D.C.: US Department of Education.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academy Press.

National Science Foundation. (2006). Math and science partnership program: Strengthening America by advancing academic achievement in mathematics and science [Electronic Version]. Retrieved June 11, 2008, from http://nsf.gov/pubs/2005/nsf05069/nsf05069.pdf

No Child Left Behind Act, PL 107-110 (2001).

Panizzon, D., & Pegg, J. (2008). Assessment practices: Empowering mathematics and science teachers in rural secondary schools to enhance student learning. *International Journal of Science and Mathematics Education, 6*(2), 417-436.

Piburn, M., & Sawada, D. (2000). *Reformed teaching observation protocol (RTOP): Reference manual* (ACEPT Technical Report No. IN00-3)

Popham, J. W. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Porter, A. C., Blank, R. K., Smithson, J. L., & Osthoff, E. (2005). Place-based randomized trials to test the effects on instruction practices of a mathematics/science professional development program for teachers. *The Annals of the American Academy, 599*, 147-175.

Radford, D. L. (1998). Transferring theory into practice: A model for professional development for science education reform. *Journal of Research in Science Teaching, 35*(1), 73-88.

Resnick, M. (2004). *The educated student: Defining and advancing student achievement.* Alexandria, VA: National School Boards Association.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. (2003). The impact of individual teachers on student achievement: Evidence from panel data. *American Economics Review, 94*(2), 247-252.

Rogers, P. J. (2000). *Causal models in program theory evaluation.* San Francisco: Jossey-Bass.

Ross, J. A., & Bruce, C. D. (2008). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education, 23*, 146-159.

Rowan, B. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools.* Paper presented at the American Educational Research Association.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70*(4), 256-284.

Sawada, D., Piburn, M., Judson, E., Turley, J., Falconer, K., Benford, R., et al. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *Measuring Reform Practices, 102*(6), 245-253.

Sawada, D., Piburn, M., Turley, J., Falconer, K., Benford, R., Bloom, I., et al. (2000). *Reformed Teaching Observational Protocol (RTOP) Training Guide* (ACEPT Technical Report No. IN00-2). Tempe, AZ.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Shotsberger, P. G. (1999). The INSTRUCT project: Web professional development for mathematics teachers. *Journal of Computers in Mathematics and Science Teaching, 18*(1), 49-60.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(9), 963-980.

Swackhamer, L. E., Koellner, K., Basile, C., & Kimbrough, D. (in press). Increasing the self-efficacy of inservice teachers through content knowledge. *Teacher Education Quarterly.*

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469-479.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations. A research report prepared for the U.S. Department of Education*: Center for the Study of Teaching and Policy, University of Washington.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* (Issues & Answers Report, REL 2007 - No. 033). Washington, DC: U.S. Department of Educationa, Institute of Education Sciences.

Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D. R., Dennison, A., Triezenberg, H., et al. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71*(1), 75-94.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343-351.

APPENDIX

A Hypothetical Example of Data Aggregation

As a hypothetical example (see Figure 2), teacher Jane Smith began her participation in the professional development treatment package in the summer of 2005. During the 2003-04 and 2004-05 school years she taught 7th grade at Jones Middle School, and during both of these years she taught four sections of mathematics (two pre-algebra and two algebra) and two sections of earth science. For Jane Smith, her pre-treatment index year was the 2004-05 school year since it was the last academic year before she started the professional development treatment, and the spring 2005 CSAP test data represent the first of two elements in the equation that producing the pre-treatment normalized gain scores for her students.

Jane Smith had 25 students in each of her four mathematics classes in 2004-05, for a total of 100 students. While some of her earth science students were also in her mathematics classes, only those students in her mathematics classes were included in the analysis. Assuming no absences or attrition in the test-taking process, all 100 of her mathematics students took the 7th grade CSAP mathematics test in March 2005 and each received a total scale score that fell within a range of 280 to 860. These 100 students' scale scores, however, were then adjusted to accommodate the fact that these students might have been unusually skilled or unskilled when compared to post-treatment classes – hence the need to create gain scores.

Archival data were used to find March 2004 6th grade CSAP mathematics total scale scores for each of these 100 pre-treatment index year students. Due to student movement in and out of district during this two-year process, it turns out in this hypothetical example that only 80 of Jane Smith's 100 7th grade mathematics students actually were enrolled in her district in the 6th grade and took the 6th grade CSAP mathematics test in the spring of 2004. Hence, it was only on these 80 common students that scale scores were calculated and the difference generated from the gain score subtraction process created.

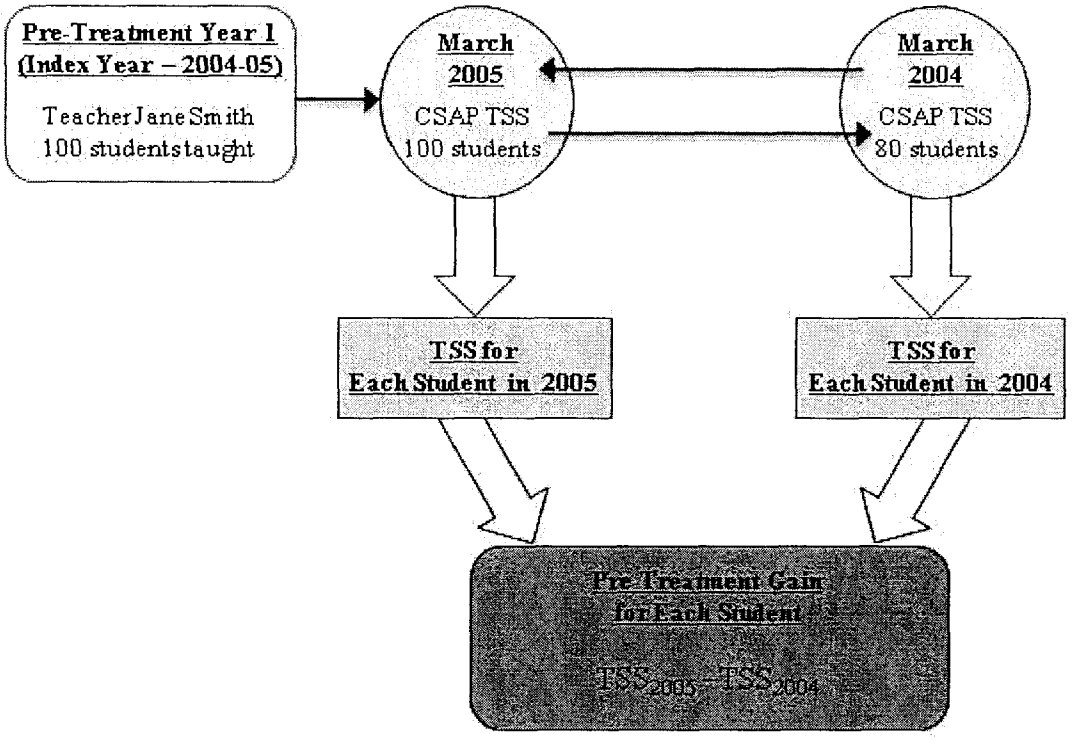


Figure A1: Data collection sequence for pre-treatment gain scores