

A CASE STUDY OF A SCHOOL DISTRICT NEW PRINCIPAL MENTORING PROGRAM

BY

LISA KAY HOOD

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Jennifer C. Greene, Chair
Professor Lizanne DeStefano
Associate Professor Donald G. Hackmann
Professor Katherine Ryan

Abstract

In an effort to learn more about evaluation theory and practice, particularly the tensions and benefits of integrating two different evaluation approaches, this study was a critical reflection of the issues and tensions related to blending program theory evaluation and constructivist evaluation into one evaluation approach that I labeled constructivist program theory evaluation. A past program evaluation of a school district's new principal mentoring program served as the "case" that underwent a reflective metaevaluation. This metaevaluation study explored two questions. First, how did integrating constructivist and program theory evaluation affect the quality of the evaluation? Second, how did this integration of two evaluation theories affect the utility of the evaluation? There were three primary data sources used to answer those research questions: an analysis of entries from a reflection journal that was kept during the time of the evaluation, interview responses from a metaevaluation panel that consisted of program evaluators and mentoring program administrators, and a comparative analysis using five program evaluations.

The findings from this analysis suggested that the strengths of the evaluation were that it provided good contextual understanding of the program, fairly represented multiple stakeholders, and that the program theory structure aided the readers' understanding of the program and led to new insights about how programs might be improved. As for weaknesses of the evaluation, the study found that the contextual factors surrounding the mentoring program were not explored as broadly as they should have, evidence of causality was not presented, and the easy convergence of multiple stakeholder program theories suggested a possible bias. The thematic analysis found three themes that highlight the importance of context in program evaluation, the difficulty of representing program complexity, and the importance of identifying clear criteria of merit and worth and clarifying their sources.

Acknowledgements

They say it takes a village to raise a child, and in this case, it took a village to earn this Ph.D. I have many people to acknowledge and thank not only for logistical support, but also for the moral and intellectual support necessary to complete such an endeavor.

First, my appreciation goes out to those educators who participated in the initial program evaluation in a school district who were not named but nonetheless were gracious to open up about their work and experiences in their schools. Second, I thank my colleagues in the program evaluation profession and the mentoring program administrators who gave of their time and expertise to participate in the metaevaluation.

I also extend a thank you to my committee members. First, to Dr. Jennifer Greene, for being my support throughout my graduate career and modeling the type of person that I hope to be someday: intelligent, kind, and generous with her time and extensive knowledge and wisdom in our field. Thank you for making a graduate student feel like a rock star every time I left your office. To my other committee members, Drs. Katherine Ryan, Lizanne DeStefano and Don Hackmann, it was a pleasure to work with you and learn from you throughout these years. I look forward to continuing to work with you in future professional endeavors.

To my Illinois State University work family, a profound thank you for your continued support not only for giving me the time when I needed to finish this dissertation, but also for never giving up on me and for having more confidence in my abilities than I think I deserve. Especially to Erika Hunt, I love you like a sister, but am so thankful that I never have to hear—“did you work on your dissertation?”—utter from your lips. Thank you for your constant pushing and for never letting me give up.

And, finally, to my family, extended and immediate—my sister, brother- and sister-in-law, aunts, uncles, and cousins. I am always in constant awe of your support for my accomplishments and I hope that I rise up to your high expectations of me. To my husband, Mark, thank you for the over 25 years of support as I have worked toward this dream, often at the expense of spending time with you. I know that this has added a stress to our marriage, but thank you for bearing with me and for being so proud of my accomplishments. To my children, Ashley Tara, and Ian, who for the first time in their lives will not have to compete with classwork or a dissertation for their mother's time. I hope that you now see that regardless of your circumstances and no matter what life may throw at you, if you have a dream you can accomplish it with perseverance and hard work. For you and for my nieces and nephews, and any children that may come into my life hereafter, all I ask as you go through life, is that you find your passion, dream your dreams, and dedicate your life to following through until you finally make those dreams come true.

Finally, we often owe our accomplishments to those who came before us, and I have been most fortunate in having a strong foundation to stand on. First, I cannot forget the profound impact that my grandparents have had on my life and my love of learning. Family lore has it that my Opa Peavler gave me one of my first beloved books that I carried with me everywhere. Every time I walked in the library stacks at the University of Illinois, I thought of my Opa Peavler and how I learned my love of reading from him and how that love for books has led to so much joy in my life. I also thank my Oma Peavler, for being so confident in me. When I was with her and talked about my studies, she was incredulous of the places I traveled and the people with whom I worked. She always made me feel so special. To my Oma and Opa Ehmen, words cannot express how much I wish you were still here to share in this accomplishment with me. A few years ago, I

found a graduation card my grandparents gave me from my 8th grade graduation. At the bottom under their signatures, my Oma wrote, “Today junior high, tomorrow the University of Illinois.” I hope that somewhere up there in heaven they give you the news that I did it. I finally made it.

Finally, I am blessed to have not only one set of parents who have supported me but two, one by birth and another by marriage. To my mother- and father-in law, Barb and Gene Hood, I thank you for your continuing support throughout these years, not only your support to me, but mostly in your support to Mark and my children. To my Dad, you may not remember or realize the excitement and gratitude for all those times when I was a girl and we used to take Sunday drives through the University of Illinois campus after Sunday dinner. Those drives nurtured my dream to attend the University of Illinois. I also thank you for the many, many times that you were a surrogate parent to my daughters and helping Mom take care of them when I was in class or studying.

And, last but not least, my gratitude knows no bounds when it comes to my mom. Without her, I truly would not have been able to earn this accomplishment. Over the years, watching you as a mother, grandmother, and daughter, I have been humbled by your selflessness. You have been a second mother to my girls anytime that we needed you and without complaints. For the countless loads of laundry; for the countless dinners; for stepping in as chauffeur; for attending sports events; and for cleaning my house before the county public health office condemned it; for the encouragement; and for the undying, unconditional love that only a mother and daughter can share—to you Mom, I thank you most of all.

Table of Contents

Chapter 1: Overview of the Study.....	1
Chapter 2: Literature Review	10
Chapter 3: Evaluation of the Clearlake Novice Principal Mentoring Program.....	38
Chapter 4: Metaevaluation Study Design.....	69
Chapter 5: Results of the Metaevaluation.....	91
Chapter 6: Lessons Learned and Implications.....	133
References.....	150
Appendix A Survey of Mentored Principals, Fall 2006: Illinois Mentoring Programs.....	162
Appendix B Interview Protocol for Clearlake District/Program Administrators.....	167
Appendix C Interview Protocol for Clearlake Principal Protégés.....	170
Appendix D Interview Protocol for Clearlake Mentors.....	174
Appendix E Self-Reflection Protocol.....	177
Appendix F Program Evaluators Interview Guide.....	180
Appendix G Mentoring Program Administrators Interview Guide.....	183
Appendix H Description of Evaluations Used for Comparative Analysis.....	187
Appendix I Comparative Analysis Protocol.....	190
Appendix J Criteria Used to Judge the Quality of the District Mentoring Program.....	193

Chapter 1

Overview of the Study

In fall 2006, I evaluated a novice principal mentoring program that is offered by a school district in Illinois. This mentoring program is geared toward all of the district's new principals in their first two years in the position. To evaluate this program, I integrated constructivist evaluation and program theory evaluation (PTE) approaches to guide the design of the study, which I labeled constructivist program theory evaluation (C-PTE) for this dissertation study. I chose these two evaluation theories because the constructivist paradigm oriented me toward the beliefs, assumptions, and values of stakeholders to help me understand how the program was implemented and the value district personnel placed on the program. Mentoring programs are reliant on the quality of the experiences and relationship between mentor and protégé, and the constructivist qualitative framework allowed me to explore these relationships to understand the program.

The PTE approach focused the evaluation on uncovering underlying theories of the mentoring program that stakeholders held in terms of identifying the problem the mentoring program was trying to solve, describing how the program was implemented, and identifying the effects the program was having on principals, their schools, their students, and the district. Incorporating PTE helped the evaluation go further than just describing the stories of stakeholders, by constructing knowledge that could be applied to other mentoring programs.

In the evaluation field, most of the published PTE evaluations largely are experimental studies to test whether a program works as intended or not. During this district mentoring program evaluation, I implemented a qualitatively dominant constructivist PTE. Howe (2004) described this type of evaluation as "mixed methods interpretivism," (p. 54) in which the

qualitative methods favor the voices of the field over quantitative assessments and representations of social phenomenon. This type of evaluation, I believed, better fit the relational nature of mentoring relationships and the indirect effects that principals have on students. Hence, the purpose of this dissertation study was to analyze the challenges and benefits of using C-PTE to conduct a high quality evaluation in terms of constructing in-depth knowledge of the evaluand and identifying a link between program participation and outcomes. In addition, this dissertation study examined C-PTE's usefulness in providing information for administrators of mentoring programs to improve their programs.

Program Theory Evaluation

In a 1987 issue of *New Directions in Evaluation* devoted to theory-driven evaluation, several definitions of program theory were brought forth by the contributors. Bickman (1987), issue editor, stated that program theory is a "construction of a plausible and sensible model of how a program is supposed to work" (p. 5). More explicitly, Conrad and Miller (1987) wrote, "program philosophy is a system of theories and values that defines the structure, population, process, and outcomes of a program" (p. 22). This definition implies that program philosophy is developed or is implicit before the development and delivery of the program and guides how the program reacts to and meets clients' needs. Sheier (1987) highlighted the causal model in programs, writing that

Program theory refers to a systematic delineation of cause-and-effect relationships.

Program theory is one or more sets of causal assumptions that provide the rationale for a program intervention. Such theory might be based simply on commonsense notions about the way the relevant actors behave. Preferably, it is derived from empirically based scientific study of the content domain for the program components being developed. (p.

61)

In one of his seminal works on PTE, Chen (1990) defined program theory as a “specification of what must be done to achieve the desired goals, what other important aspects may also be anticipated, and how these goals and impacts would be generated” (p. 43). Weiss (2000) defined program theory as “links between what programs assume their activities are accomplishing and what actually happens at each small step along the way” (p. 35).

Program theory evaluation uses the program theory to guide the evaluation design, methods, analyses, and recommendations. In PTE, the evaluator attempts to uncover the theories that guide program implementation and identify the mechanisms that lead to program outcomes (intended or unintended). The evaluators identify the theories that describe how the program “should” be implemented and its intended outcomes, as well as theories actually in use. Evaluators also may identify the theories that link the program’s processes and its outcomes. Evaluators who use PTE attempt to explain whether the program’s success or failure is due to faulty program theories that guide program implementation, or is due to program administrators’ or practitioners’ deviation in implementing the program.

Supporters of PTE assert that PTE helps the audience understand why programs succeed or fail and explicates the links between program processes and its effects (Rogers, Petrosino, Huebner, & Hacsı, 2000). In addition, evaluators can help the program’s efficiency by identifying the essential processes that lead to program success and identifying those non-essential processes that do not contribute to the success of the program. Identifying essential program processes allows the program administrators and designers to make the best use of their time and resources (Birckmayer & Weiss, 2000; Weiss, 1997).

Critics of PTE assert that evaluators who use PTE may make an erroneous assumption that programs are grounded in a validated theory that can be identified prior to the evaluation, which may not always be true (Stufflebeam, 2001a). In addition, testing theories is time and resource intensive. Scriven (1998) criticized PTE for relying too heavily on the goals of the program to guide the evaluation, which may lead evaluators to miss unintended program effects. Furthermore, PTE often treats programs as fixed units when, in reality, programs evolve in response to stakeholders and the environment. Thus, critics say evaluators who implemented a time intensive and expensive PTE could end up with a program theory that is not an accurate representation of the program (Davidson, 2006). Lastly, Schwandt (1993) criticized PTE for using sophisticated statistical modeling techniques to develop and test program theories that evaluate the program theory rather than evaluate the program.

Significance of the Study

Because the evaluation field is relatively young, it is populated with evaluation theories that are not yet validated (King, 2003). Consequently, for over 20 years evaluators have called for empirical studies of evaluation practice to compare the utility and feasibility of the different evaluation theories and models and study the utility of prescriptive theories (Henry & Mark, 2003; King, 2003; Shadish, 1998; Smith, 1993; Stufflebeam, 2001b; Worthen, 2001). The evaluation field has evolved over the years responding to the political and contextual needs of programs and stakeholders. Studies of evaluation practices have investigated stakeholder involvement, stakeholder use of evaluation recommendations, client/stakeholders relationships, and evaluation's influence on policy decisions (Brandon & Fukunaga, 2014; Johnson, Greenesid, Toal, King, Lawrenz, & Volkov, 2009; Smith, 1993).

Smith (1993) stated that empirical knowledge based on evaluation practice is vital in the development of relevant and useful descriptive evaluation theories that describe when an evaluation theory works best, how it functions best, and what it can hope to accomplish and under what conditions. Smith (1993) also asserted that studies of evaluation practice should assess the utility of prescriptive evaluation theories that describe what evaluators should do to conduct a “proper” evaluation. A prescriptive evaluation describes the function and role of evaluation, the evaluation questions that should be asked, the evaluation design and implementation, and utilization of results (Shadish, Cook, & Leviton, 1991). Program theory evaluation is one such prescriptive theory that I chose to critically analyze in this study.

In this study, I explored the issues and challenges of using program theory evaluation (PTE) within a constructivist evaluation framework as a model for providing in-depth understanding of the underlying assumptions and structure of the district mentoring program and making causal linkages between the program and outcomes. A significant portion of the literature on PTE is approached from a postpositivist paradigm using quantitative statistical and modeling techniques to test program theories and program effects. Dahler-Larson (2001) argued for integrating PTE and constructivism. Adding constructivist concepts to PTE would underscore the interdependent relationship between a program and its context. Moreover, Dahler-Larson (2001) asserted that constructivist concepts would bring in multiple stakeholders’ experiences and values and document the evolutionary nature of programs as they change in response to the changing needs of stakeholders and the environment. By integrating constructivist and PTE approaches, I responded to Dahler-Larson’s call to build a bridge between constructivism and PTE to represent stakeholders’ beliefs and values; describe how the program evolved since its inception; and to identify the interdependent relationship between the program,

the district, and the attitudes and beliefs of program stakeholders that influence the program's effectiveness.

Research Questions

This dissertation is a reflective metaevaluation of a constructivist program theory evaluation of a district-sponsored new principal mentoring program in Illinois. The questions that guided the study aim to inform the evaluation community about the challenges and benefits of employing constructivist program theory evaluation to evaluate programs and provide quality and useful information to program stakeholders. This dissertation addressed the following primary research questions:

1. How did integrating elements of constructivist and program theory evaluation affect the quality of the evaluation in terms of:
 - a. Constructing knowledge about how the district mentoring program works;
 - b. Representing the multiple theories of stakeholders;
 - c. Describing the relationship between the program and its context; and
 - d. Identifying the program's effects on principals, schools, and students?
2. What is the utility of a constructivist program theory evaluation for learning about principal mentoring programs in terms of:
 - a. Helping program administrators understand mentoring programs and make changes or improvements to their own programs;
 - b. Informing the leadership field about how mentoring programs work and the effects they have on principals and their schools?
3. How does this evaluation compare to a sample of constructivist, program theory and realist evaluations in quality and utility? Did those evaluations meet the quality and

utility criteria to a greater or lesser extent than mine, or did they face many of the same or additional challenges?

The first question sought to provide information on the quality of C-PTE from the perspectives of evaluators in the field, including my own reflective and comparative analysis of the evaluation study. The evaluation study was metaevaluated by an external panel of evaluators. My reflective analyses and the metaevaluation review panel used a blend of criteria to metaevaluate the quality of the constructivist PTE of the novice principal mentoring program including: constructivist criteria (Guba & Lincoln, 1989), the Joint Standards for the Evaluation of Educational Evaluation criteria (1994), constructivist grounded theory (Charmaz, 2006), and realist theory criteria (Julnes & Mark, 1998; Mark, Henry, & Julnes, 1998).

During the C-PTE of the mentoring program, I anticipated the generation of multiple program theories. To answer this first research question, I reflected on the challenges of reconciling the underlying values and experiences of multiple stakeholders. I reflected on the challenge of representing the mentoring program's theories that capture stakeholders' values and experiences and the contingent relationship between the mentoring program and the sponsoring school district.

The second question explored the usefulness of C-PTE for learning about principal mentoring programs. In addition to convening evaluators on my metaevaluation panel, I also recruited mentoring program administrators to review the evaluation report and respond to a set of questions regarding the usefulness of the information using constructivist criteria (Guba & Lincoln, 1989) and utility criteria from the Joint Committee Standards (1994). The findings of this study will inform the evaluation literature by describing the results of my attempt to bridge

constructivism and PTE in terms of the quality and utility of the information gleaned from such an evaluation.

The third question consisted of a comparison of my C-PTE against a sample of evaluations to examine whether these evaluations were similar or different to mine in meeting the rigors of the evaluation criteria used in this metaevaluation. These evaluations were conducted in the PTE, interpretivist PTE, realist, and constructivist models of evaluation. With this analysis, I hoped to explore whether there were certain themes in the challenges and benefits of using these models in creating high quality and useful evaluations.

Limitations of the Study

One of the significant aspects of this study is that it examines the connection between theory and practice using one specific evaluation context as the case under study. However, because it is one case, this study is limited in its generalizability as a critique of C-PTE. This study provides depth through my reflections of the case, the review by the metaevaluation panel, and the comparisons to other evaluations. However, this metaevaluation is my own; the reflections are my own, as well as the interpretations of the reactions of the panel and practitioners. Therefore, the findings of this study may not generalize to other forms of constructivist or program theory evaluations that differ from the model used in this study.

In addition, another limitation of this study is the passage of time between the dates of the evaluation of the new principal mentoring program, this metaevaluation study and the publishing of the results. Every effort has been made to remain current with the literature and policies that are related to principal mentoring programs and the field of program evaluation. Where the literature is directly tied to the data collection and analyses of either the program evaluation of the mentoring program or the metaevaluation study (e.g., mentoring program literature used for

the a prior program theory development, program evaluation literature tied to the quality and utility criteria), I retained the original literature that I used to guide my data collection and analyses efforts. The last chapter contains an update on the policies and context that relate to new principal mentoring programs and new developments regarding potential data sources that could contribute to the evaluation of new principal mentoring programs in Illinois.

Study Overview

The second chapter of this dissertation outlines the program theory evaluation literature highlighting the concepts of key PTE theorists as they defined PTE, its purpose and promise for more comprehensive and informative evaluations, as well as exemplar evaluations that model the concepts of PTE. The third chapter provides a description of the C-PTE evaluation of one school district's novice principal mentoring program that served as the case study for this dissertation. This chapter includes a summary of the mentoring program and the evaluation design. The fourth chapter presents the design of this dissertation study, the questions under exploration and the methods and criteria were used to metaevaluate the C-PTE of a district's new principal mentoring program. Chapters five and six present the findings of the study, as well as the summary of themes and implications for the evaluation field.

Chapter 2

Literature Review

This literature review highlights the core components of program theory evaluation (PTE). This review focused on PTE because PTE was the more dominant evaluation model used for the C-PTE of the new principal mentoring program that is the focus of this metaevaluation. In lieu of introducing the elements of constructivist evaluation that were integrated in the evaluation, I describe in succeeding chapters the elements of constructivist evaluation that were integrated with PTE with appropriate citations of the literature to support their use in the evaluation. The review begins with the impetus for developing this evaluation model and continues with a presentation of the key definitions and concepts of the model, including the general definitions of program theory as well as the different conceptions of program theory. After this presentation, various models of program theory evaluation are described including graphic representations of the models and examples of the models in practice. In addition, the review includes a description of both the advantages and criticisms of PTE. Finally, the review concludes with a recommendation to integrate PTE and constructivism to enable evaluators to make both value claims and causal claims that more accurately represent the fluid nature and contextual realities of programs.

Theory in Evaluation

Program theory evaluation (PTE) grew out of evaluators' and stakeholders' disenchantment with methods-driven experimental evaluations and outcomes evaluations that provided information on the effects of programs but could not provide the evaluators and stakeholders with the information about why or how programs succeeded or failed (Chen, 1990; Donaldson, 2003; Funnell & Rogers, 2011; Rogers, 2003). These evaluations typically were

labeled black box evaluations, in which the evaluation focused on the outcomes of the program, but did not collect adequate information on the program processes or intervening or causal mechanisms that occurred within the black box that produced the outcomes (Funnell & Rogers, 2011). Without comprehensive information about program contexts and processes, local stakeholders and policymakers did not have valid information on which to base decisions on program improvements, make policy decisions such as funding or resource allocation, or replicate programs across settings (Chen, 1990; Donaldson, 2003). Thus, stakeholders and policymakers often made hit-or-miss improvements or decisions based on anecdotal information or their own beliefs about what went wrong with the program. Or conversely, they replicated what they thought went right with the program with little empirical evidence to substantiate these beliefs.

Over the years, various evaluators have called program theory evaluation by many names: theory-driven or theory-based evaluation (Chen, 1989; Chen 1990; Donaldson, 2003; Weiss, 1997; Weiss, 1998), program theory evaluation (Davidson, 2000; Funnell & Rogers, 2011; Rogers, 2003), and logic modeling (Adler, 2002; Frechtling, 2007; Kellogg Foundation, 2004; Kent, Donaldson, Wyrick, & Smith, 2000). For the purpose of this review, I used the term program theory evaluation (PTE) to refer to this model of evaluation. I chose this nomenclature versus the alternatives such as theory-driven or theory-based because, similar to Davidson and Rogers, I believe it most closely situates the evaluation in the program, its contexts, practices, and outcomes. In contrast, the terms “theory-driven” or “theory-based” can imply the use of grand social science theories, which were not used in this study.

Program Theory Evaluation Conceptual Framework

Program theory evaluation uses the program’s own theory, or underlying assumptions

and beliefs to guide the evaluation design, the development of the evaluation questions, and the selection of methods to test the program theory (Bickman, 1996; Chen, 1990; Weiss, 1997).

Chen (1990) asserts that programs are purposeful and organized efforts to provide a service or solve a problem. Theories are assumptions or principles that explain or guide social actions. PTE is an approach to evaluation that uses the theory that explains how a program solves a targeted problem to guide the evaluation design, data collection activities, analysis, and recommendations. In PTE, the evaluator attempts to uncover the explicit and implicit theories that guide program implementation and identify those mechanisms that lead to program outcomes (intended or unintended). Essentially, evaluators uncover the causal model(s) that underlies programs and then test its efficacy. This type of evaluation is theory-driven rather than method-driven; thus, evaluators have a smorgasbord of evaluation tools and a choice of quantitative, qualitative, or mixed methods designs, rather than anchoring the evaluation in a particular methodology. Typically, the evaluator portrays the program theory in a graphic or pictorial form with lines and arrows that delineate cause-and-effect relationships.

Many evaluation theorists have contributed to the formulation of PTE, most notably Huey-Tsyh Chen. Chen (1990) provided a comprehensive conceptual framework for PTE in which he categorized different types of program theories into two overarching categories. First, Chen (1990) described the normative theories (i.e., prescriptive theories) that describe how the program “should” work in terms of the treatment, implementation processes, and outcomes. The normative theories identify program goals or outcomes and guides the design and implementation of the treatment.

Chen (1990) labeled the second type of theories, causative theories (i.e., descriptive theories), that identify the underlying causal mechanisms. These mechanisms establish the

linkages between the program's implementation processes and outcomes. The causative theory is not a description of how the program should be, but of how it is. Bickman (1987) and Chen (1990) stated that PTEs test for program theory failure/success or implementation failure/success. More specifically, the evaluation seeks to explain whether the success or failure of a program is due to erroneous beliefs or assumptions that guide program implementation (i.e., theory failure) or due to a lack of fidelity in implementing the program as intended in the program's theory (i.e., implementation failure). This information is useful for program stakeholders because it tells them whether they need to re-think the basic assumptions of the program, or whether they need to identify where their implementation broke down and make corrections in delivery.

Similar to Chen, Weiss (1998) delineated PTE into two types: implementation theory and program theory. Implementation theory describes the program activities, more specifically the intensity, fidelity, and quality of program activities. Program theory focuses on the mechanisms that mediate between program activities and outcomes. Weiss (1998) further emphasized that the treatment is not the mediating mechanism. Rather, the mediator is the response that the treatment generates that leads to the outcomes of program participation. The combination of the two types of theories results in a program's theories of change (Weiss, 1998). Often it is the work of the evaluation to test program theory and implementation theory to determine if the failure of the program is due to program theory failure (i.e., the failure of the program is due to erroneous assumptions or a faulty social science theory), or if the failure is due to implementation failure (i.e., the program is not delivered as intended by the program theory) (Chen, 1990; Sheirer, 1987; Weiss, 1998).

In his conceptualization of PTE, Lipsey (1993) identified six elements that comprise

program theory: problem definition, mediating processes, expected output, exogenous factors, critical inputs, and implementation issues. The problem definition draws the boundaries under which the program operates in terms of the condition it is treating, the targeted populations, and the context. The mediating processes refer to causative mechanisms that lead to the transformation and resulting outcomes. The expected outputs describe the effects that program stakeholders expect as a result of the program. The exogenous factors are the other factors found in the program's context or environment that potentially affect the intervention. Critical inputs are those required processes that the program has to supply in order to achieve the intended effects. Finally, implementation issues are those issues that are related to the intervention's delivery system. These issues refer to the resources needed to deliver program services as intended including both material and human resources.

To identify how and why a program works is one of the key concepts of PTE. This identification is accomplished by identifying and testing the intervening mechanisms that contribute to the likelihood that the program's intended outcomes will be achieved. As noted above, Chen (1990), Lipsey (1993), and Weiss (1998) refer to the causal or intervening mechanisms that are the program's causal agents. The intervening mechanism is not the treatment or intervention but the participants' response to the intervention. Mark, Hofmann, and Reichardt (1992) and Petrosino (2000) refer to these intervening mechanisms as mediators and moderators. According to these evaluators, mediators are the causal links between x and y ; they explain the cause-and-effect relationship between the program and its effects. For instance, Donaldson and Gooler (2003) studied the effect of the Winning New Jobs program that provided job search training to un- and under-employed California residents with the goal that these program participants would become reemployed and enjoy improved mental health. The

mediators that were hypothesized to lead to these ultimate outcomes were increased job search confidence, improved job search skills, and more problem-solving strategies.

Moderators, on the other hand, are those mechanisms that influence the size of the program's effects. The relationship between x and y (i.e., the program and its effects) varies as a function of z (i.e., moderators) (Mark, et. al., 1992). For example, an evaluator studying the effects of a substance abuse program may analyze whether or not the program is more effective for males or females. The participants' gender would be the moderating variable in the program. However, Petrosino (2000) found that including moderator variables in an *a priori* program theory model is rare, and that most analyses of moderator effects occurs *post hoc*. Petrosino stated that including an analysis of moderators at the end of an evaluation was a problem, and that more attention should be given to identifying potential moderator variables when developing the program theory, rather than leaving them as an afterthought.

Purpose of program theory evaluation. Chen and Rossi (1992) asserted that the primary purpose of PTE is for “policy decision-making and practical applications” (p. 5). Weiss (1997) declared that PTE is warranted when program stakeholders want to know why and how their program has succeeded or failed and if only one or two relatively long-term effects of moderate complexity are under investigation. In addition, PTE may clarify the issues surrounding a program if previous studies show inconsistent results. PTE also facilitates cost- and time-efficient replication of programs when it identifies those essential intervening mechanisms that promote successful implementation and effects.

According to PTE advocates, PTE can be used by program developers and staffers to gain more in-depth understanding of their programs. Often, they have a vague notion of what makes the program succeed or fail, but the program's theories may be ambiguous or unstated

(Bickman, 1987; Chen, 1990; Chen & Rossi, 1992; Donaldson, 2003; Weiss, 1998). However, because a program often is not built on or guided by one theory, it is not the evaluator's goal to uncover the "best" or "correct" program theory but to uncover a pertinent program theory (or set of theories) that can guide the evaluation efforts (Bickman, 1987).

Program theory sources. The origin or sources of program theory should be examined. Evaluators have suggested that program theories typically are derived from two sources: social science theories, and the beliefs and assumptions of stakeholders (Bickman, 1987; Chen, 1990; Chen & Rossi, 1992; Weiss, 2000). Therefore, in developing program theory, evaluators have several options. First, some evaluators may choose to refer to the social science (or educational) literature that provides research findings on particular issues and typically report on potential interventions that alleviate the problem. The advantage of using social science theories is that there is usually a systematic, cumulative literature base on which to draw. However, the disadvantage is that often theories can be abstract and disconnected from practice (Weiss, 2000).

Second, to develop local program theory evaluators can use interviews, document analysis, and surveys to collect program stakeholders' values, beliefs, and implicit and explicit assumptions about the effects of their actions on program participants'/clients' knowledge, skills, and behavior. Stakeholder values are especially important because they determine what problems and populations should be targeted with intervention efforts and guide the development and implementation of the intervention (Chen, 1990; Weiss, 2000). Therefore, it is crucial to uncover stakeholder values to examine their influence on program implementation and effects, as well as disagreements or inconsistencies among program stakeholders regarding how the program should be and is being implemented and its effects. Thus, Chen (1990) advocated that evaluators use both social science theories and stakeholder perspectives in a PTE.

Prioritizing program theories for study. Evaluators using PTE may find that program stakeholders report multiple theories about the work they do and its effects. Because most evaluators have limited time and resources, they must choose which theories and causal links to explore. Weiss (2000) identified four criteria for selecting theories. First, evaluators should consider the beliefs of program stakeholders. Which program theories do stakeholders believe are having the most effects? Evaluators have a wide repertoire of methods to collect these perspectives including: interviews and focus groups (Donaldson & Gooler, 2003; Green & McAllister, 2002), document analysis (Adler, 2002; Donaldson & Gooler, 2003; Green & McAllister, 2002), surveys (Conrad & Miller, 1987), and outcome pattern matching and concept mapping (Calvin, Streeter, Kim, & Tripodi, 2011; Rosas, 2005; Trochim, 1989; Yampolskaya, Nesman, Hernandez, & Koch, 2004). Using these methods, evaluators reveal stakeholders' most salient beliefs and assumptions in order to develop program theories and select the theories that seem plausible for influencing program effects. These program theories are used to guide the evaluation.

Second, the evaluators should examine the plausibility of the theories. In other words, can the program actually deliver the activities it intended to deliver and will the target population react to the program's inputs as expected? To answer these questions, Weiss (2000) recommended examining the program's itemized budget or other program documents to see where time and money are being spent, and examine what resources are being provided to implement the program. If it does not appear that the program is being implemented according to theory, then there is no reason to follow that particular theory. For example, if one theory for an educational program that intends to raise student achievement in math stipulates that individualized instruction for students will meet the learning needs of students but teachers are

not given the time to individualize the instruction for each student, then it would be futile to examine this particular program theory because teachers are not being given the time to implement the theory of individualized instruction.

Third, evaluators may choose to examine a particular program theory because little is known about that particular causal mechanism and its effects. Therefore, such evaluations may add to the literature. Finally, evaluators may choose to evaluate a particular program theory because it is essential to the success of the program and without this particular theory, the program would not achieve positive effects (Weiss, 2000). Using the example of the math program in the previous paragraph, if stakeholders believe that individualized instruction is the theory that is most essential to program success and achieving improved math achievement, then that program theory should be tested. The next section discusses various ways of representing program theory in PTE.

Program Theory Representations

This section outlines the most common program theory representations typically seen in the evaluation field. These representations include representations of program theory evaluation, realist evaluation, and logic models.

Traditional program theory representations. Program theories often are presented in a graphic or diagram. These may be in the form of a flow chart, a series of boxes, a concept map, or some other type of diagrammatic representation that typically uses lines and arrows to represent the causal relationships between program inputs and processes and both immediate and long-term outcomes. Typically causal models are variable-oriented, representing the program treatment, intervening mechanisms, and outcomes as variables, and using the model to illustrate the covariate relationships among variables (Lipsey & Pollard, 1989). Some can be rather simple,

presenting the relationship between one intervening mechanism and its outcome (Figure 1).

Others are more complex, representing multiple mediating or moderating factors, their intended and unintended outcomes, and outside influences on both the factors and outcomes (Figure 2).

The last type of causal model depicts the theory in a series of boxes to lay out the inputs, processes, products, and outcomes (Funnell & Rogers, 2011; Rogers, 2003; Rogers, Petrosino, Huebner, & Hacsı, 2000) (Figure 3).

Figure 1

Simple Causal Model with One Intervening Mechanism

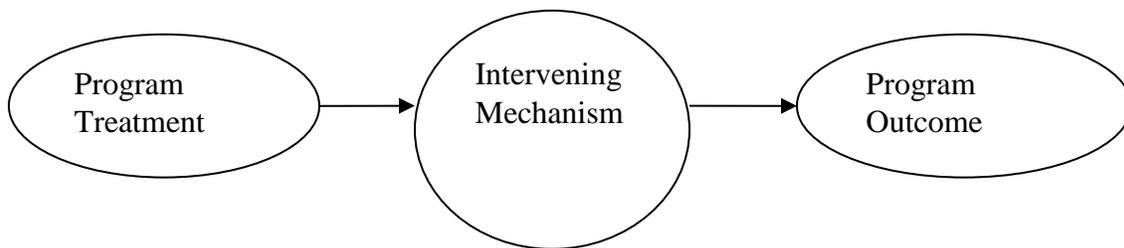


Figure 2

Complex Causal Model with Multiple Mechanisms and Extraneous Factors

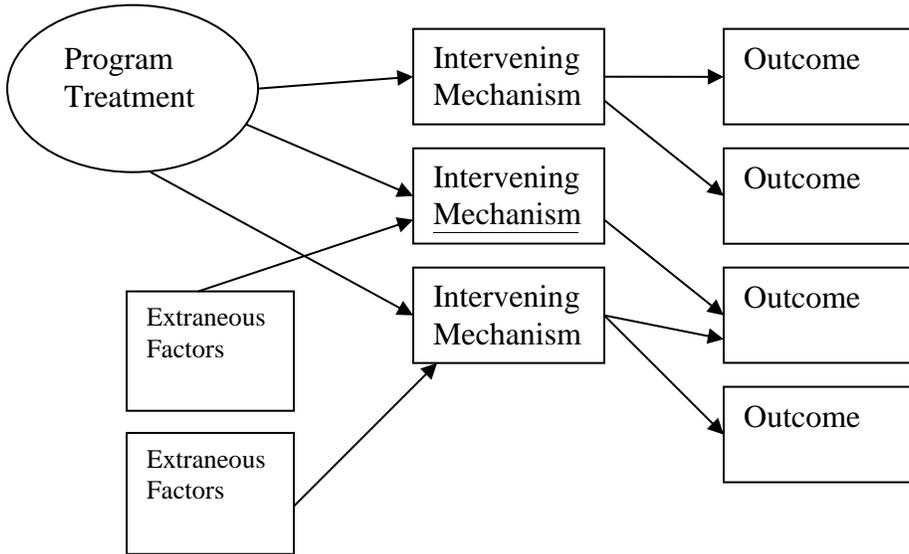
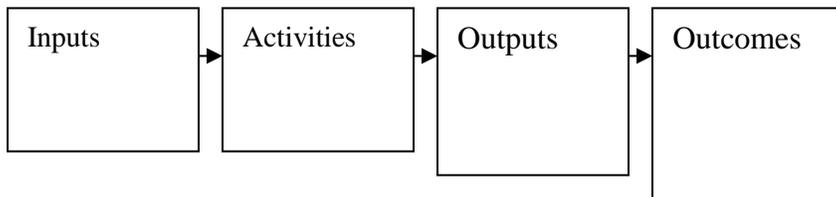


Figure 3

Simple Causal Model



Program theories also may be represented as several parallel causal models in which the program uses multiple, complementary mechanisms to achieve the ultimate outcomes (Rogers, 2003). For instance, Rogers (2003) used the example of teacher home visits to students' homes, intended to improve student learning by engaging parents in the learning process. The mechanisms identified in the program theory were teachers spending time listening to the

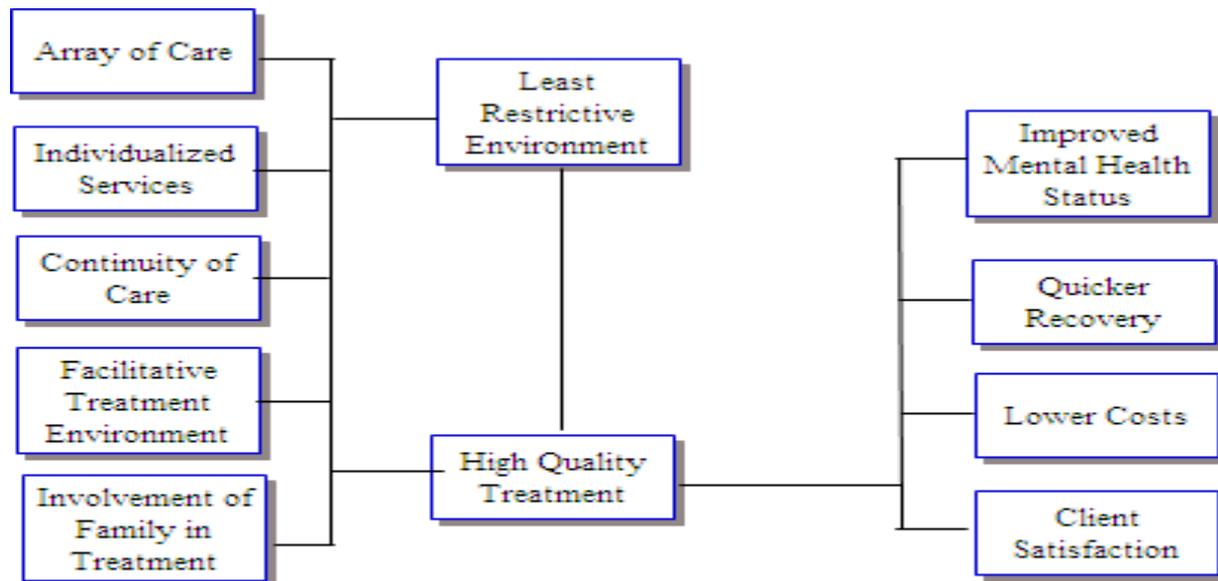
student, teachers seeing students' home environments, and teachers and parents spending time with students. These parallel and complementary activities led to students believing that the teachers care about them, teachers seeing the barriers to student learning in their homes, and parents learning to trust teachers. The short-term outcomes are students cooperating with teachers in the classroom, teachers adapting instruction to students' interests and learning needs, and parents supporting teachers' efforts. These short-term outcomes then lead to the ultimate outcome of the program, which is improved student achievement.

Bickman (1996) wrote about the application of PTE in the evaluation of the Ft. Bragg Child and Adolescent Mental Health Demonstration project, a managed care project that provided mental health care and substance abuse treatment to children with such problems. The purpose of this evaluation was to assess whether a continuum of care to children was an effective means of providing care that led to improved mental health status, a quicker recovery time, lower costs, and higher client satisfaction. The program theory was developed using data collected in interviews with program developers and administrators after the evaluation was funded but two months before delivery of services. This preliminary model consisted of five program inputs, two intervening mechanisms, and four ultimate outcomes (Figure 4). This preliminary model guided the beginning of the evaluation process, but document analysis and interviews with program administrators and staffers led the team to develop a more detailed model of the program's theory. Although the ultimate outcomes did not change, the concepts of program inputs became more complex and were categorized according to program intake, patient assessment, and treatment. In addition, program outcomes were divided into proximal outcomes and intermediate outcomes that ultimately lead to the intended final long-term outcomes. For instance, at program intake, the evaluators identified four program inputs: a single point of entry, no required co-

payment or deductible, prompt intake, and increased public information. These inputs led to two proximal outcomes. First, patients had better access to mental health treatment and had increased awareness of treatment opportunities. These two outcomes in turn led to two intermediate outcomes in which the program saw more clients and these clients expressed higher satisfaction with the program's services. At this point this state of the program theory influenced the second stage, the client assessment phase. The inputs and outcomes in the assessment phase then influenced the program theory components in the treatment phase, which ultimately influenced the final program outcomes (Bickman, 1996)¹.

Figure 4

Reproduction of the Fort Bragg Child and Adolescent Demonstration Preliminary Program Theory Model (Bickman, 1996, p.113).



Realist evaluation. Pawson and Tilley (1997) stress the important role of context in their conceptual framework of realist evaluation. Program theories in this evaluation approach are

¹ Note: the final program theory was too complex to recreate in this document.

conceptualized in context-mechanism-outcome (CMO) configurations that highlight the influence of context on the intervening mechanisms and outcomes, recognizing that some program interventions only are effective in certain contexts. To represent these models, Pawson and Tilley construct matrices of context-mechanism-outcome configurations in which the text describes the mechanisms and outcomes and the context in which mechanisms are effective. Essentially, the model asserts that Context + Mechanism = Outcome.

Pawson and Tilley (1997) provide the final CMO configuration from an evaluation of the Safer Cities program, a British crime reduction program for neighborhoods with high crime rates. Table 1 offers an excerpt of the CMO developed from the Safer Cities evaluation findings. This model illustrates the typical layout of the CMO model.

Table 1

CMO from the Realist Evaluation of the Safer Cities Program (Pawson & Tilley, 1997)

Context	Mechanism	Outcome
Central government removed discretionary power of local authorities	1. Safer Cities resented as central government intrusion. Efforts to appropriate available grants to make up shortfall in revenue	1. Safer Cities operates as 'suspicious'
Central government reduced local authority revenue raising opportunities	2. Safer Cities becomes part of the local institutional landscape, builds local capacity for crime prevention and is found useful for money provided, advice given, and demonstration that agencies working together could achieve more than working in isolation	2. Safer Cities acts as an 'honest broker'

Compared to other models, the CMO model (Pawson & Tilley, 1997), more explicitly addresses the effects of context on program implementation and outcomes. Pawson and Tilley recognized that the causal value of programs lies in the contingent relationship between contexts and the mechanism by which the program hopes to achieve its outcomes. The example in Table 1 provides the context by briefly describing the actions of the British central government and provides two possible mechanisms by which the context may be interpreted by the local communities. Whichever interpretation the communities adopt ends in one of the two possible outcomes. For instance, the central government developed a program to alleviate high crime rates in identified communities, which essentially took control from the local governing bodies and authorities to allocate resources toward fighting crime. The local communities could interpret the government as either an intruder or a benefactor. To understand the CMO configuration of a particular program, Pawson and Tilley suggested that evaluators should examine the different combinations of beliefs, actions, assumptions, ideals, rules, norms, and environmental setting inherent in the program's context. In other words, evaluators should investigate the social, economic, and cultural factors in which programs are situated that serve as the backdrop for the allocation of community resources and rationale for programming that explain how and why a program works and under what circumstances.

Thus, Pawson and Tilley (1997) posed a stronger emphasis on moderators and how moderators influence program practice and outcomes, either positively or negatively. As the model suggests above, certain contexts and moderators (mechanisms) inhibit program success (mechanism 1), while others facilitate it (mechanism 2).

Logic models. The last models presented are the logic models. Some evaluators use the terms PTE and logic models interchangeably. However, others identify differences between logic

modeling and traditional PTEs particularly in their focus on program theory components. For instance, Cooksy, Gill, and Kelly (2001) stated that one of the major differences between logic models and PTE is that program theories are rooted in social science theories, while logic models are rooted in the specific practice of programs. Others cite the differences in representing program theories. Although program theories can be narratives or graphics representing program theory components such as activities, moderators, intervening mechanisms, and outcomes, the logic model typically presents the program theory model as a flow chart or a series of boxes that displays the logical sequence of events (Cooksy et al., 2001; Frechtling, 2007; Kellogg Foundation, 2004; Plantz, Greenway, & Hendricks, 1997).

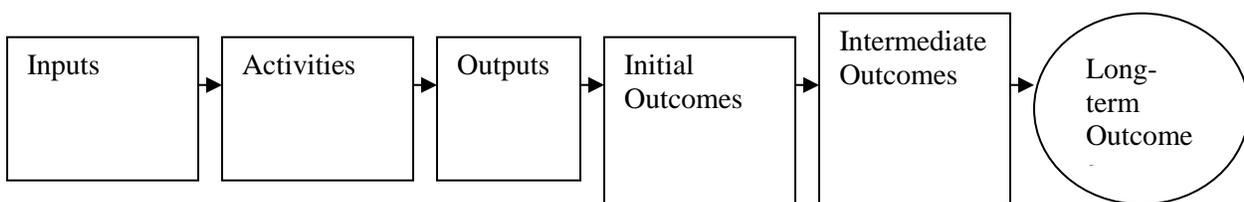
Because of the increasing call by program funders for measuring outcomes, logic models have become increasingly popular and more funding agencies are requiring grantees to submit program logic models with applications and use them to monitor and evaluate program implementation and impact (Kaplan & Garrett, 2004; Kellogg Foundation, 2004). The logic models presented in this review were developed by the United Way of America and the Kellogg Foundation for use by community-based initiatives to develop, monitor, and evaluate their human service programs.

The United Way of America promoted its logic model framework through the Outcome Measurement Resource Network, which called its model the “Program Outcome Model” (United Way of America, 1999). The development of the outcome model was in response to a need to measure and report program outcomes rather than listing only program outputs. Program outputs are the immediate record of program activities, for example, the number of units delivered or the number of people served, while program outcomes are changes in program participants’ behavior, or their response to the program intervention. The United Way model represented the

program logic as a chain of program inputs, activities, outputs, and multiple levels of outcomes from initial to intermediate to longer-term outcomes (Plantz, Greenway, & Hendricks, 1997). (Figure 5). The authors referred to these as “if-then chains” such that if the program provides “x” resources, then it can provide “y” activities. If “y” activities occur, then “z” outputs are produced, and so forth. In each box, the program identifies and lists program processes and outcomes. Evident in this model was a heavy emphasis on outcomes. The literature shows that the United Way promoted the collection of data on outcomes rather than outputs as a performance measurement system for program managers. United Way indicated that agencies that collected outcomes data reported several benefits: providing a focus for program administrators and staffers, serving as a barometer to gauge success and make necessary adjustments, motivating program staff who see tangible results of their work, recruiting volunteers, and raising the image of the program in the community as a successful tool for social change (Hendricks, Plantz, & Pritchard, 2008).

Figure 5

United Way Program Outcome Model



In the W.K. Kellogg Foundation’s (2004) “Logic Model Development Guide,” Kellogg defined logic models for Kellogg grantees as “a picture of how your organization does its work—the theory and assumptions underlying the program. A program logic model links

outcomes (both short- and long-term) with program activities/processes and the theoretical assumptions/principles of the program” (p. iii). The Kellogg framework is very similar to the United Way’s and contains the program’s resources/inputs, activities, outputs, outcomes, and effects.

For both the United Way and Kellogg logic model frameworks, the resources/inputs refer to the program human and financial resources that are needed to operate the program. Activities refer to those program processes, tools, and events that program staffers use to implement the program. The remaining components of the model represent the program’s intended effects. There are multiple levels of outcomes. First, the program’s outputs are the direct program products. Outcomes represent the changes in participants’ behavior, knowledge, and skills. The logic model differentiates between short-term and long-term outcomes, initial and intermediate outcomes (United Way of America, 1999; Kellogg Foundation, 2004). According to the Kellogg Foundation (2004), short-term outcomes should be attained between one to three years, and long-term outcomes in four to six years. Both the United Way and The Kellogg Foundation models identified and measured the ultimate outcomes (both intended and unintended) and captured these in the long-term outcomes and impacts columns. According to Kellogg, program impacts should be achieved within seven to 10 years.

The National Dropout Prevention Center/Network produced a logic model that highlights nine strategies for building community collaborations between the Collaborative Truancy Prevention Task Force and their local community partners to reduce the truancy rates in their communities (Cash & Cloud-Duttweiler, 2005). The purpose of this logic model was to guide local initiatives in building these collaborative partnerships and accomplishing the work of the initiative, thereby leading to the long-term outcome of increased school attendance and student

achievement. This model included columns for the inputs that feed into the programs outputs, which in this case are the program activities, short-term and long-term outcomes. This model also included a list of assumptions that undergird the logic of the model and will lead to increased student attendance and achievement. These assumptions included total and long-term commitment by all stakeholders, valid and reliable data for decision-making and annual evaluation to guide strategies. This model also identified the following external factors (which the Kellogg and United Way models do not): family and school factors, economic influences, student variables, and the community context (Cash & Cloud-Duttweiler, 2005). It is unclear in the report how the local initiatives were using the model to guide and evaluate their programs.

The last two sections of this review present the advantages of implementing PTE as well as some critiques of PTE.

Advantages of Program Theory Evaluation

According to Bickman (1989) there are 10 benefits of using program theory in evaluation. Focusing evaluations on program theory will: contribute to social science knowledge, assist policymakers, discriminate between theory failure and program failure, identify the problem and target group, provide a description of program implementation, uncover unintended effects, specify intervening variables, improve formative use of evaluation, clarify measurement issues, and improve consensus formation (p. 387). Essentially, evaluators use program theory to identify not only if a program works but also how and why (Rogers, 2000a).

Furthermore, advocates of PTE assert that formulating program theories and using them to guide evaluation has certain advantages. PTE advocates claim that PTE leads to information that helps evaluators and stakeholders understand why programs work or do not work. Because it examines the links between the program and its effects, PTE can make more credible causal

attributions to the program (Rogers et al., 2000). PTE identifies the essential mechanisms that lead to positive effects and potentially identifies unnecessary program components that do not affect program effects, thus allowing program designers, administrators, and staff to maximize their time and resources on those essential causal mechanisms (Birckmayer & Weiss, 2000; Weiss, 1997). Identifying these essential mechanisms also allows program designers to replicate programs and adapt these mechanisms to their own environments and target populations, while omitting the non-essential components from the program (Weiss, 1997).

Using PTE, evaluators also can identify intervening mechanisms that at the time of the evaluation do not appear to affect participants, but appear to be suffering from delayed feedback and exhibit the potential for long-term impact (Birckmayer & Weiss, 2000; Donaldson, 2003; Weiss, 1997). PTE can contribute to the cumulative knowledge base and/or social science theories that inform the development and implementation of programs that serve society's needs (Bickman, 1987; Birckmayer & Weiss, 2000; Chen, 1990). During PTE, program practitioners undergo a process of intense scrutiny and reflection on their program processes, assumptions, and goals which may benefit program developers and practitioners in the following ways. PTE allows them to:

- examine the feasibility of their program logic and determine if their ideas and intended outcomes are coherently linked,
- examine the coherence of their program logic across practitioners in the program and develop a shared set of assumptions, and
- examine the effectiveness of their program logic and make improvements to the logic and practice to increase program effectiveness. (Weiss, 1997)

For evaluators, PTE guides them to focus the evaluation on key questions related to the

program's theories of change, helping them to make crucial choices to focus their limited time and resources (Weiss, 1997). Donaldson (2003) proposed that PTE forces evaluators to think and reflect *before* they act. Otherwise, evaluators often fall back on their preferred methods rather than reflect on the match between evaluation questions and best methods for studying the program. Evaluators also may identify short-term outcomes that may lead to long-term impacts (Weiss, 1997). In addition, PTE helps evaluators to develop recommendations for future program improvements that are more useful and relevant for program stakeholders; thus PTE is an appropriate process for formative evaluation (Rogers et al., 2000; Weiss, 1997). Finally, Donaldson (2003) advocated for PTE because he believes that PTE can lead to greater stakeholder inclusion to formulate and test program theories, more so than other traditional, methods-driven approaches. The next section describes the critics' responses to PTE.

Critiques of Program Theory Evaluation

Program theory evaluation is not without its critics—one who outright rejected PTE as a viable approach to evaluation (Stufflebeam, 2001a), and others who recognize weaknesses in the approach but acknowledged the benefits of PTE with modifications (Dahler-Larson, 2001; Davidson, 2006; Scriven, 1998). Stufflebeam (2001a) based his rejection of PTE on several factors. First, PTE assumes that programs are grounded in sound and validated theory when in fact many programs may not be. When there is no explicit theory guiding the program, then the evaluator is placed in the role of program theory developer. The theory then needs to be validated and tested, which may be problematic as the evaluator assumes the developer role from program administrators. In addition, validating and testing theories is time and resource intensive which may strain the evaluator and program's resources. Finally, often PTE involves developing the program's logic model, which Stufflebeam (2001a) stated is "armchair theorizing" (p. 38) by

the evaluator and program stakeholders who presumably know how and why a program is expected to achieve its desired outcomes. Theory developed in such a way does not meet the standards of sound theory. In addition, often PTE treats programs as static entities, when in truth, the program may have evolved over time and the original program theory may no longer be relevant to the program. Thus, the end result of such an evaluation could be an invalidated theory that was expensively formulated and is a misrepresentation of the program. Thus, Stufflebeam (2001a) concluded that “overall, there really is not much to recommend theory-based program evaluation, since doing it right is usually not feasible and since failed or misrepresented attempts can be highly counterproductive” (p. 39).

Davidson (2006) summarized common criticisms to PTE categorizing the criticisms into four “baggages:” crystal ball baggage, goal-based baggage, stakeholder involvement baggage, and hypothesis-testing baggage. Stufflebeam’s (2001a) criticisms would be categorized in the crystal ball baggage in which PTE assumes that a program theory and the conditional relationships, variables, and intended effects can be identified a priori to the evaluation, when often these components of program theory emerge during an evaluation. The second baggage, goal-based, criticizes PTE for relying too heavily on the stated goals of the program as a significant source of program theory. Scriven (1991, 1998) argued that evaluators should study whatever effects a program generates independent of the intended goals. Knowing the goals may in fact create tunnel vision that focuses the evaluator on goal attainment and misses unintended program side effects. At most Scriven (1998) recommended that evaluators develop a program logic model to examine what components may be faulty in the case of program failure. Stakeholder involvement baggage includes those criticisms that PTE is an exclusively participatory approach involving stakeholders in program theory development to the exclusion of

other sources and methods of developing program theory. However, contextual or contractual barriers may inhibit stakeholder involvement in the evaluation (Davidson, 2006). Finally, the last category of baggage, hypothesis-testing, alludes to the common practice of statistically testing hypotheses about program processes and outcomes that A leads to B. As mentioned in earlier sections of this chapter, many of the program theory evaluations have involved sophisticated statistical modeling techniques to create and test patterns of program variables. However, the focus then becomes on evaluating the program theory rather than the program and what is less evident is whether or not the program effects made a meaningful, or practical, difference in anyone's lives (Schwandt, 1993).

In a critique of PTE, Funnell & Rogers (2011) said that “when the same approach to program theory is used for all types of interventions and all types of purposes, the risk is that the interventions will be distorted to fit into a preconceived format” (p. xxi). In response, these evaluators proposed an approach to PTE that they called “Purposeful Program Theory” (p. xxi). With this approach, the evaluator carefully assesses the context of both the program theory and the program or intervention to consider the purpose of the program theory in terms of who will use it and how, and the nature and context in which the program or intervention is situated. They urge evaluators to not think about PTE as a “one-size-fits all approach” (Funnell & Rogers, p. xxi).

A constructivist critique of program theory evaluation. Traditional PTE grew out of a technical social engineering paradigm that had previously focused its research efforts in experimental designs, which ultimately earned the name of “black box evaluations.” Analyzing PTE from a constructivist viewpoint, Dahler-Larson (2001) criticized PTE for not adequately addressing the social constructive features of programs, particularly the contextual features in

which programs operate. Program contexts are fluid and adaptive, rather than static. Of particular importance are those contextual features that serve as moderators and can influence the effectiveness of the program, either positively or negatively. The crux of Dahler-Larson's critique against PTE is the unfortunate lack of attention on moderators.

Dahler-Larson (2001) recognized the importance of moderators as powerful contextual features that can strengthen, weaken, or even nullify the program theory. Chen (1990) used the example of a job skills program that teaches participants job skills with the intended outcome that program participants will obtain jobs after finishing. Chen identified the market-need of the taught job skills as an important moderator—if the job skills are not relevant to the job market, employers will not hire program completers, thus, nullifying the program theory. Pawson and Tilley (1997), in the realist evaluation CMO explicitly recognized the contingent relationship between the program's contextual features and the mechanisms of change that achieve program outcomes. However, what these evaluation theorists ignored is that moderators are fluid and adaptive social constructions that often change in the interactions between program stakeholders and the program. Dahler-Larson (2001) illustrated this point with three types of programs: tragic, magic, and competing programs.

Tragic programs are characterized by a self-defeating program theory that creates the conditions that facilitate its failure (Dahler-Larson, 2001). Using Chen's job skills program as an example, a tragic program is one in which the message implicit in program activities (e.g., teaching job skills and aiding in job searches) may signal to program participants their inability to control their own lives. Thus, the program participants may not be successful in obtaining jobs because they lack the self-confidence in their abilities to succeed in that job. Thus, the program interacted negatively with the moderator of participants' self-confidence and damaged the

program's opportunity for successful outcomes.

In contrast, magic programs are self-fulfilling programs in that they create the conditions that guarantee their success. Following up with the job skills program example, a magic program might establish a successful public relations campaign that creates buy-in and enthusiasm in the community, which in turn inspires community members increasingly to hire program completers. Thus, the program interacted positively with the moderator of community support, and it is this interaction of conditions and the program that facilitated program success, not the program inputs and activities alone. Finally, competing programs are characterized by competing program theories that work against each other in program operation. Finishing with the job skills program example, a competing program would be one in which one group of instructors may believe that being successful in the workplace requires adequate knowledge and technical skills, while other instructors may believe that success in the workplace requires the dispositions (e.g., attitudes and motivation) that define a good work ethic. Program participants receive conflicting messages throughout their program and may become confused of what is expected from them. Thus, program success may be hampered by these competing stakeholder program theories.

As a remedy, Dahler-Larson (2001) suggested building a bridge between constructivism and PTE recommending that evaluators map out the commitments and values stakeholders place on program components and contexts. Thus, a PTE would be characterized by representations of multiple program theories, and the contingent relationships between the program and the moderators. Evaluators also should expand their inventory of causal designs to more accurately reflect the interactions between programs and their contexts, making it explicit where the program may knowingly or unknowingly affect its context, which in turn affects program implementation and effectiveness. In essence, in traditional PTE representations, the arrows that

point from moderators to programs are typically unidirectional pointing from moderator to program. A constructivist approach would represent this relationship with bidirectional arrows pointing to and from moderators and programs indicating the interdependent relationship between a program and its context. In advocating for conducting PTE through a constructivist lens to improve evaluation results, Dahler-Larson made this claim:

Theory-based evaluation with a constructivist inclination helps evaluators more clearly to identify the values people connect to outcomes, programmes, programme theories, and moderators. Subsequently, evaluators can produce both value statements and causal statements about program theories with a higher degree of precision. (p. 347)

Summary

This literature review highlighted the core features of program theory evaluation. The review began with a short history of PTE and its various definitions. Essentially, PTE is focused on uncovering the beliefs, assumptions, and values of program actors that influence program implementation and effectiveness. It is concerned with describing the primary activities and structure of a program, the context within which the program is implemented, and the effects (immediate, short-term, and long-term) that the program has on participants within the context of the program. Program theory can have various sources such as social science literature and literature in the substantive field, as well as stakeholder beliefs and values. The review summarized the common models of PTE including realist evaluation and logic models.

Finally, the review presented both the advantages of PTE and its critics. Advocates of PTE claim that this evaluation model provides comprehensive and targeted information about how and why a program is successful or a failure which program designers and administrators can use to improve and replicate programs. In addition, PTE advocates claim that PTE adds to

the cumulative knowledge on programs.

In contrast, critics claim that PTE can be resource intensive and that the program theories that are developed may not be an accurate representation of the program's present reality. Furthermore, PTE assumes static program theory(ies) and does not take into account the fluid and contingent relationships between the program and its context which may positively or negatively influence the program's success. Critics also claim that PTE assumes that programs are based on a well-grounded and validated program theory that can be identified a priori to the evaluation. Finally, PTE practitioners also are criticized for relying too heavily on statistical analysis to evaluate the components of the program theory which ends in an evaluation of the program theory rather than the program.

A C-PTE will address these critics in the following ways. First, constructivist evaluation is an attempt to understand program stakeholders' meanings and experiences with the evaluand and reveal the salient contextual features of the environment in which programs interact. Therefore, a C-PTE acknowledges that programs are not static, nor are their contexts or the program participants. Thus, the evaluator documents the evolution of the program and allows program theory to emerge during the course of the evaluation, acknowledging that an evaluation may generate multiple program theories. In addition, constructivist evaluators inquire into the interplay between the program's context and program implementation and outcomes, recognizing that the relationship between context and programs is interdependent (Dahler-Larson, 2001). A C-PTE places primacy on qualitative methods that elicit stakeholders' beliefs, values, and assumptions as related to the program through dialogue. The evaluator is looking for patterns of meaning that represent the stakeholders' views of how the program works and what it means to them (Howe, 2004). However, PTE adds a focus on framing the data on stakeholder

meanings, beliefs, and experiences collected in a constructivist PTE in terms of developing these data into a program theory that explains the effective mechanisms of a program and into a map of causal linkages between program practices and effects.

The next chapter provides a description of the program theory evaluation of the district-led new principal mentoring program within a constructivist paradigm to address those criticisms of PTE established by Dahler-Larson (2001). This evaluation serves as the case study of PTE to analyze the challenges and issues of using PTE as the guiding framework for conducting quality and useful evaluations that provides valuable information for evaluators in their work and for practitioners to develop and improve programs.

Chapter 3

Evaluation of the Clearlake Novice Principal Mentoring Program

This dissertation study is a metaevaluation of an evaluation of a novice principal mentoring program. This chapter describes the evaluation that served as the “case” that was the focus of this metaevaluation dissertation study.

The novice principal mentoring program implemented² by the Clearlake School District was the focus of the program evaluation. While serving in another administrative role prior to being named the superintendent, Dr. Smithson proposed the mentoring program for new principals in a grant proposal to a funding agency that created an initiative focused on improving educational leadership in states and districts across the country. The proposal was successful, and the Clearlake School District received significant funds to implement leadership initiatives such as an administrative intern program, school improvement coaches, a cohort education administration program in partnership with an Illinois university, and professional development with new principals including the novice mentoring program.

Beginning in 2004, all new principals (called protégés in the program) in the district were required to participate in the new principal mentoring program. Upon their hiring, each new principal was paired up with a retired or practicing Clearlake school district principal who served as the protégé’s mentor for his/her first and second years. Protégés and their mentors were required to meet at least once a week each month of the school year for two years. At the beginning of the school year, the protégé and his or her mentor developed the protégé’s goals for the year, and the mentor supported the protégé’s efforts in reaching those goals throughout the year. Mentors described themselves as critical friends who prompt their protégés to reflect on

² Pseudonyms were used to protect the confidentiality of the school district, administrators, and other participants in the evaluation.

their practice and talk through strategies for dealing with a particular leadership issues in their schools. The mentors did not see their role as someone to tell the principals what to do, but rather to guide their protégés in problem solving and reflection on their practice. Personnel in the district's human resources office oversaw the mentoring program with the goal toward developing effective and adaptive instructional leaders for the district's K-12 education system. District leaders expect principals to serve as change agents in their schools. Change agents are adaptive leaders who modify their leadership practices to meet the needs of their community, their school, their teachers, and their students. Adaptive leaders challenge the status quo, leading faculty and staff to make the changes necessary to improve student performance. From its inception in 2004 through the end of this evaluation which was conducted in spring 2007, the program had mentored 23 principals and assistant principals.

Origin of Interest

Since 2000, The Wallace Foundation has provided grants to states and districts to improve school leadership capacity. In 2001, Illinois State University's Center for the Study of Education Policy along with organizations in 15 other states were awarded funding by The Wallace Foundation to examine the conditions and policies that affect education leadership in each state and to make recommendations to strengthen school leadership that would lead to improvements in student learning. In Illinois, this project is called the Illinois State Action for Education Leadership Project³ (IL-SAELP).

After conducting a comprehensive study of education policies between July 2001 and December 2003, IL-SAELP members made approximately three dozen recommendations that

³ IL-SAELP was a consortium of educational stakeholders from various organizations including higher education educational administration programs, school administrator and school board organizations, policymakers, teacher unions, local district and school leaders. The project was managed by one faculty member and three staff members at Illinois State University. This project ended in 2009.

addressed a broad spectrum of leadership concerns, such as, delineating and redefining the roles and responsibilities of the superintendent and teachers as leaders, strengthening principals as instructional leaders and refocusing principal preparation programs to prepare principals for this role, increasing diversity in Illinois school leaders, and creating more partnerships between universities and school districts. In 2004, some of the group's recommendations were put into action through legislation to revise the current certification requirements for principals and other school leaders, such as teacher leaders.

After many rounds of discussion and negotiation with various education stakeholders, this legislation, Senate Bill 860 (Public Act 094-1039), was passed in the Illinois legislature in May 2006. This new legislation created an advanced certification system that mandates participation in a one-year mentoring program for new principals (contingent on funding) that was called the Illinois New Principal Mentoring Program, an advanced endorsement for Master Principal designation after completing a Master Principal Program, alternative routes to the principalship for National Board Certified Teachers, and a teacher leader endorsement for those teachers who want to serve in leadership positions without having to leave the classroom (Illinois Public Act 094-1039, 2006).

I became interested in issues surrounding school leadership as a staff member of the Illinois State Action for Education Leadership Project (IL-SAELP) from 2005 to 2009. My work focused on school leadership has continued with additional grants received from new funders. In this role, my primary work duties have been to research the information needs of school and district administrators to be effective instructional leaders. A large focus of my work has had a birth to grade three (i.e., early childhood) focus and what educational leaders need to know in order to effectively lead programming at this level in their schools. My work has also provided

forums for networking among a consortium of principal preparation programs as they have redesigned their programs in response to new state program requirements, as well as provide staffing and technical assistance to the Illinois State Board of Education in facilitating education stakeholders to recommend new program requirements for superintendent preparation programs.

From an initial study of the preparation and professional development strategies of Illinois school leaders, as an IL-SAELP staff member, I found that mentoring was viewed as a viable strategy to provide support to new principals as they adjust to the complexity and demands of their new position. I surveyed all public school Illinois principals and superintendents regarding principals' preparation and professional development needs and found that these school leaders supported the statewide mentoring program (IL-SAELP, 2005). First, nearly half of responding principals ($n=379$) stated that their preparation for the position was irrelevant to practice and that they lacked exposure to practicing principals, real school settings, and the complexities of the job. Further, 62% of principals ($n=466$) asserted that the state should pay for and require mentoring for first-year principals (IL-SAELP, 2005). In addition, those principals who had participated in a formal mentoring program early in their careers valued their mentoring experience and were more likely to support a mandatory mentoring program than those principals who did not have this experience.

In general, superintendents also voiced their support for mentoring. Forty percent of responding superintendents ($n=154$) stated that the state should require a mentoring program for new principals. In telephone interviews with 10 Illinois superintendents, eight supported the mentoring program, and in interviews with 19 superintendents in Arkansas, Delaware, and Oregon (states that already mandate mentoring), nearly all superintendents recommended that Illinois adopt a mandatory statewide mentoring program for its new principals (IL-SAELP,

2005). Researching novice principal mentoring programs and their worth to practitioners in the field became one of my special interests as we continued to advocate for the statewide mentoring program for first-year principals.

Program Evaluation of the Clearlake Novice Principal Mentoring Program

Evaluation Study Design

This evaluation study was a constructivist program theory evaluation (C-PTE) of a district-sponsored mentoring program for new principals hired by the district. The study was a mixed methods design driven by qualitative methods that included a survey to former protégés, open-ended interviews, and document reviews. The primary purpose of the evaluation was to construct knowledge about how mentoring programs work and the effects the programs have on new principals' leadership development and practices from the viewpoints of the district/program administrators, principal protégés, and mentors. The audience for this evaluation was broad, including researchers and practitioners in the educational leadership field as well as policymakers. A secondary purpose of this evaluation was to provide information to the district/program administrators about the worth their principals and mentors placed on the mentoring program and the program effects reported by principal protégés and mentors. Program administrators could use this information to make program improvements.

Phase one: a priori program theory development. To begin the evaluation, I reviewed studies of principal mentoring programs to capture the effects mentoring programs were having for new principals and their schools. From this review, I was able to glean insights into the purposes and goals of novice principal mentoring programs, those components of mentoring programs that researchers attributed to mentoring programs' success (or failure), the potential outcomes of mentoring programs for new principals, and possible challenges or barriers to

effective program implementation. The following paragraphs summarize the literature I used to develop the a priori program theory that guided the program evaluation.

First, mentoring programs serve two important functions: to develop the novice principal as an effective school leader by supporting the new leader as he or she learns and practices new skills and knowledge, and to socialize the new leader into the role of principal (Browne-Ferrigno & Muth, 2004; Daresh, 2004). To fulfill these functions, researchers and program practitioners have identified essential components for a mentoring program. These include: mentor training in both technical (e.g., imparting experiential knowledge) and cognitive (e.g., promoting self-reflection mentoring) skills (Bloom, Castagna, Moir, & Warren, 2005; Villani, 2006; Wallace, 2007); mentor selection criteria focused on the mentors' communication, interpersonal, and leadership skills (Geismar, Morris, & Lieberman, 2000; Villani, 2006); systematic process and responsibilities for the mentoring relationship assigned to both the mentor and protégé (e.g., regular meeting times, reflection opportunities) (Villani, 2006); a mentoring relationship based on trust and respect (Zellner, Skrla, & Erlandson, 2001); and district support of the program (Villani, 2006). The Wallace Foundation (2007) issued a report on novice principal mentoring program that listed criteria for quality mentoring programs. In addition to high quality mentor training mentioned above, the list also included:

1. Data collection on program efficacy (i.e., how the program is (or is not) building leadership behaviors and dispositions)
2. At least one year of mentoring, two years is ideal
3. Funding that is sufficient to provide mentor stipends, and
4. Clear program goals that are focused on providing new principals with the knowledge and skills to provide leadership directed at improving teaching and learning.

After researching basic components or practices of a mentoring program, the literature review turned toward identifying possible outcomes of mentoring. A major focus of the studies and reports on mentoring programs has been the benefits of mentoring programs and relationships. Benefits generally have been documented for three groups: the protégés, mentors, and the schools and districts in which mentors and protégés lead. Benefits to protégés include receiving empathy, encouragement, and counseling, collaborating and sharing ideas and practices, and receiving constructive criticism on their work, identifying areas for needed improvement, becoming more goal-oriented and reflective of their knowledge and practices (Ehrich, Hansford, & Tennet, 2004; Rich & Jackson, 2005). In multiple studies of mentoring, researchers found that new school administrators generally reported increased confidence in their professional competence (Daresh, 2001, 2004; Daresh & Playko, 1993; Hansford & Ehrich, 2005; Playko 1995). In addition, researchers found that principals were able to translate leadership theories learned in their training program into practice in their schools, developed enhanced communication skills, and learned new administrative and leadership skills (Bloom et al., 2005; Browne-Ferrigno & Muth, 2004; Playko, 1995; The Wallace Foundation, 2007). Principals also reported a new sense of belonging within their professional community, because the mentoring relationship reduced their feeling of isolation and socialized them into the broader context of schooling (Browne-Ferrigno & Muth, 2004; Playko, 1995). Enomoto, Gardiner, and Grogan (2000) discovered that female school administrators who had an informal mentor advanced more quickly than those women who did not. Often, mentors serve to broaden the new principals' networks, and through these networks, new principals have opportunities for career advancement (Browne-Ferrigno & Muth, 2004; Playko, 1995).

The benefits of mentoring programs have not been limited to the protégés. In their studies, researchers also found that mentors and the school districts that provide mentoring opportunities for new school leaders reported benefits as well. Benefits for mentors include: collaborating and networking with colleagues, having time for reflection on their own professional knowledge and skills, sponsoring their own professional development and growth, experiencing personal satisfaction and reward for helping junior colleagues, being recognized as a valuable resource to teach others and enjoying being a teacher again, and feeling less isolated (Ehrich et al., 2004; Hansford & Ehrich, 2005; Playko, 1995; Villani, 2006). An evaluation of The Principals Excellence Program (PEP), a federally-funded in-service professional development program offered in the Pike County School District in Kentucky, found that mentoring principals increased the quality of the leadership capacity within their district, offered the district the opportunity to offer professional development that is tailored toward districts' needs, and changed the culture of the district and schools by forming a spirit of collegiality (Browne-Ferrigno & Muth, 2004).

While research on mentoring has broadened its focus to study the effects of programs on mentors and districts, few studies have attempted to follow the causal chain of effects to establish links between effects of the mentoring program on novice principals, their schools, and student learning despite the fact that research has shown that principals' leadership behaviors do affect students. In fact, research on the effect of principals has found that principal behaviors account for approximately 25% of school effects on student achievement (Leithwood, Louis, Anderson, & Wahlstrom, 2004). Principal effect on student learning is second only to the effect of teachers. However, researchers have asserted that principals have an indirect effect on student learning with much of their influence bearing on school and classroom conditions, which in turn affects

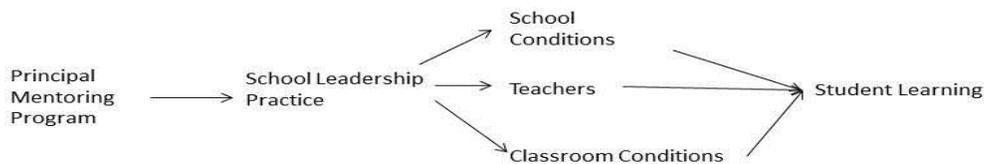
student learning (Hallenger & Heck, 1996 ; Leithwood et al., 2004; Waters et al., 2003).

Therefore, in order to be able to evaluate mentoring programs' effect on new principals and subsequently their schools and students, one would have to follow a causal chain of effects.

Figure 6 is a model of a how a mentoring program affects principals and the subsequent effects on schools, classrooms, and students. I relied on this model as a template for thinking about how a novice principal mentoring program could affect not only the novice principal but also his/her schools and students.

Figure 6

A Model of the Effects of a Mentoring Program on Principals, Schools, and Students



Adapted from Leithwood et al., 2004, p. 18

Finally, in addition to the benefits of mentoring programs, researchers have identified potential limitations or unintended consequences of participating in mentoring programs. Playko (1995) identified two types of barriers that limit the effects of mentoring programs. These barriers included those that can occur in the planning processes and those that arise as problems

in implementation. For example, in the program planning stage, the program may not receive the full support of the district, which can lead to a program that is not fully conceptualized beyond the pairing up of protégés with “buddies” in the school district. Often lack of district support led to less commitment by mentors and protégés and/or a relationship that primarily is focused on the technical skills of the principalship that new principals need to survive. In contrast, an effective mentoring program focuses on promoting “leadership not survivorship” blending both the technical and instructional leadership skills required to successfully lead schools for improvement (Bloom et al., 2005; Playko, 1995; Villani, 2006).

A lack of district commitment may lead to the second type of barrier that affects program implementation including: lack of time, burden on protégés and mentors, poor protégé and mentor matches, and poor mentor training (Playko, 1995). As a result of weak commitment, district leaders may not reserve the time needed for mentor and protégé to meet. The most common limitation that participants have reported is a lack of time to fully implement the program or to fully engage in the mentoring process. Program participants also have complained about poor matches between mentor and protégé based on professional expertise and/or personality, and that some mentors are out of touch with today’s real educational problems and did not receive the training needed to fully understand the goals of the mentoring program (Ehrich et al., 2004; Hansford & Ehrich, 2005; Playko, 1995). Often, program administrators make the erroneous assumption that any experienced administrator has the qualifications to mentor a junior colleague.

Through this literature review I constructed an a priori program theory of how and why mentoring programs may affect new principal professional development. In narrative form, the a priori program theory was this: a new principal is paired up with a mentor, and during the

mentoring process, the mentor and this principal protégé build a relationship based on trust in which the principal protégé can expose his or her weaknesses and fears without fear of retribution or disrespect. The principal protégé exposes the problems he or she is facing in leading his or her school, and together the mentor and protégé develop strategies for addressing these problems. From this mentoring relationship, the principal protégé develops and refines his/her leadership practices, gains self-confidence in his/her leadership abilities, and feels more comfortable in the role of school leader. As a result, the principal will be more effective in building a positive school culture and improving the quality of teaching in the school which will result in more favorable student outcomes (e.g., higher attendance rates, less discipline problems, and improved student achievement). This a priori program theory guided the program evaluation in my design of the evaluation questions, instrument development, analysis, and reporting. I chose to develop this a priori theory as a way to organize my thoughts and inquiry and methodological decision-making for the evaluation of the mentoring program.

Evaluation questions. The evaluation study was guided by the following evaluation questions:

1. What is the underlying rationale or logic that guided the design/adoption and implementation of the mentoring program? In other words, why do program administrators and other stakeholders believe that these mentoring programs would aid the professional development of new principals?
2. What are the outcomes of the mentoring program in this district? What effects is the program having on:
 - a. Protégés and novice school leaders? (e.g., improved leadership practices, increased confidence and adjustment to the position)

- b. Schools?
 - i. Effects on the school culture (e.g., collaboration among the principal and faculty, focus on learning, high expectations of students and adults)
 - ii. Effects on teachers and instruction in the classrooms (e.g., increased teacher satisfaction and morale, improved teaching practices)
 - iii. Effects on students (e.g., student learning on achievement tests and other assessments, attendance, graduation rates)
- 3. Why does this mentoring program work (or not work)?
 - a. What are the primary causal mechanisms that contribute to program success (or failure)?
 - b. What are those essential components or conditions that must be in place to achieve successful outcomes?
 - c. What barriers exist that impede successful program implementation and inhibit positive program outcomes?
 - d. Are there any unintended consequences that result from this program?

Question one was answered with interview data from program administrators and the superintendent, principal protégés, and mentors, which resulted in the formulation of program theories for each group of stakeholders participating in the evaluation. These program theories represented the groups' beliefs about how and why the program was effective in building new principals' leadership capacity. The second question guided the collection of outcomes data from the program participant survey, interviews, teacher surveys, and school performance data. These data were used to test the effectiveness of the mentoring program. In other words, these data provided insight regarding whether or not the mentoring program was effecting leadership

development and practice, and whether this development translated into improvements in the protégés and their schools. Finally, the third question addressed the mediating and moderating factors that further explain how and why the program is or is not effective. This question was explored through the former participant survey and interview data with all three groups in the district case study evaluation. The interview questions focused on the contextual features inherent in the program (e.g., school demographics, gender, race/ethnicity) that may have contributed to the size of the effect the mentoring program may or may not have had on leaders' development and school improvements. The remainder of this chapter includes a more detailed description of the study.

Phase two: Survey to former mentoring program participants. The a priori program theory developed as a result of my literature search on mentoring programs was reflected in a survey distributed in November 2006 to former mentoring program protégés from new principal mentoring programs across Illinois. This was a survey that was distributed to larger group of principal protégés that included protégés from the Clearlake School District as well as several other programs that had been offered in the state at that time. The purpose of this survey was to collect feedback from the protégés about their perspectives on the level of effects the mentoring program had on their leadership practices and the strengths and weaknesses of the program. A copy of this survey can be found in Appendix A. The Clearlake mentoring program participants were part of this sample, and there was a choice in the demographic section of the survey that asked survey respondents to identify the program in which they participated so that I could pull their data for a program specific analysis for my C-PTE of the Clearlake new principal mentoring program. Administrators of the Illinois mentoring programs provided participant lists and contact information. The survey was sent to 230 mentoring program participants and 70

former and current protégés returned completed surveys. Clearlake School District mentoring program administrators provided participant lists of all participants who were mentored since the program's inception, excluding those who were in their first year of the principalship (or other school leadership position) and in their first year of the mentoring program. Surveys were mailed to 20 former mentoring program protégés from this list provided by the Clearlake mentoring program administrators. At the time of this evaluation, there were 34 principals in the school district. However, only new principals in their first year of serving in the role of a principal or assistant principal were required to participate in the district's new principal mentoring program. Therefore, the survey only went to those principals who had participated in the mentoring program. The majority of these former protégés were working as principals or assistant principals in schools throughout the district. I chose this sample because I wanted protégés who had at least one full year of leadership experience to reflect on the effect of the mentoring program, and the survey went to all persons on this list. The results of the survey for all protégés who responded were used to write a white paper about mentoring programs in Illinois that my colleagues and I shared with educational leadership stakeholders across Illinois. The Clearlake protégés' responses were part of that dataset. For my evaluation of the Clearlake mentoring program, I disaggregated the Clearlake protégés' data from the larger data set.

The results of this survey informed the development of the instruments for the third phase of the evaluation such as the interview protocols and teacher surveys. With the survey results, I looked at the effects that this mentoring program was having on school leaders on a general level, and the survey responses were also compared against the a priori program theory. During the next phase of the study, I moved to examining the particular experiences of program administrators and participants within the program.

Phase three: Local program theory development and testing. The third phase of the study consisted of collecting more in-depth perspectives through interviews from Clearlake program mentoring program stakeholders including the program administrators, mentors, and former and current protégés, who were working in Clearlake schools as principals. This phase of the evaluation was conducted in April and May 2007.

I began this phase of the evaluation by interviewing the two mentoring program administrators and the superintendent. The program administrators are district administrators in the central office who have as part of their duties the professional development of school leaders and teachers. The evaluation continued with successive interviews with five protégés and four of their mentors⁴. The superintendent and one of the program administrators developed this program and have intimate knowledge of the original rationale and assumptions that guided the mentoring program design.

The interviews with the program administrators, superintendent, principal protégés, and mentors were open-ended, using an interview protocol to guide the discussion. The purpose of these interviews was to uncover the administrators' perspectives on how and why the program should be effective in producing highly qualified leaders for the districts' schools (i.e., the program theory). Program administrators were asked questions about the problem or conditions the program seeks to address, the primary goals of the program, and the primary components and mechanisms by which the program intends to achieve its goals. The administrators also helped me understand how the mentoring program fits within the district's overall professional development activities and the district's culture and vision of schooling. The interview protocol guide for program administrators is located in Appendix B.

⁴ One of the mentors moved out of state and had been out of touch with the district and the mentoring program for two years.

The evaluation continued by sampling from former and current novice principals (protégés) who participated or were participating in the district mentoring program at the time of this evaluation. At this point of the study, the design incorporated a series of mini-case studies using current and former protégés as the unit of analysis. A sample of five Clearlake protégés was recruited to participate in the evaluation through interviews. For each protégé, I collected data from his/her mentor. The protégés' mentors were also recruited to participate in an interview about the mentoring experience and the effects of the program on their protégé. School level data from the school's interactive report card and school improvement plan, including student performance data also were collected. I attempted to collect data on measures of school climate and teacher attitudes through a teacher survey, but the response rate was too low to be able to use the results.

Student performance data in each principal's first year was used as benchmarking data to examine whether or not the principal was positively affecting student learning in the school. I explored the general upward trajectory of student performance (e.g., student learning, student attendance, graduation rates) since the principal took his/her position as school leader of that school. Similar to the teacher survey, through the interviews with the protégés, mentors, and administrators, it became very clear that the student data would not be a good measure of the effects of the mentoring program's effect on protégés' performance. Many had been in the position for only a little over one year, and the others had only been in the position for two to three years. Through the interviews I heard about many of the complex issues that the protégés were facing in their schools (e.g., teacher resistance to the protégés' leadership style and practices, changing student demographics with more poverty and mobility). So, many of the protégés, mentors, and administrators did not feel that the protégés had been in the school long

enough to detect strong effects in their schools. Therefore, the analysis of student data was abandoned.

Sampling procedure. I worked with the two program administrators from the district's human resource department to choose the five principals (i.e., protégés) from their mentoring program participant list who served as the subjects of the mini-case studies. The sample was limited to five principals due to budget and time constraints, limiting the amount of qualitative data that could be reasonably collected and analyzed in a timely manner.

To be included in the sample for the evaluation, the protégé had to serve as a principal within the Clearlake School District with one or more years of experience as a principal. I also chose protégés of both genders, principals of color, and protégés from schools that have a significant disadvantaged student population. In addition, the protégés in the study were from all levels of schools (three elementary schools, one middle school, and one high school). It was valuable to include protégés from these groups to explore if protégés from different backgrounds experience mentoring programs in similar or dissimilar ways and if there are some mentoring practices that are more or less effective for some groups of principals than others.

In April and May 2007, at each of the school sites, I conducted interviews with the protégés to gather their perspectives on their experiences in the mentoring program. The interviews consisted of open-ended questions. The protégé interview protocol is located in Appendix C. The purpose of these interviews was to ask protégés to describe their mentoring relationship and whether or not they felt the mentoring program effected their leadership development. Primarily, these protégés were asked whether or not the mentoring program helped them develop the competencies and confidence needed to effectively administer their school and serve as instructional leaders who foster a vision of improved teaching and learning. In addition,

these protégés were asked how and why they believe the mentoring program was effectively mentoring new principals and how it was initiating them into the culture of the district. The purpose of these interviews was to develop the program theory from the protégés' perspective as a test of the a priori program theory, and then test the theory as to the effectiveness of mentoring programs to help novice principals grow into effective leaders who shape their schools into effective learning organizations for adults and students alike.

To continue the case study, each of the protégé's mentors also was interviewed, asking similar questions as those asked of protégés. The mentor interview protocol is included in Appendix D. Mentors were asked to describe their experiences as mentors in the program, reflect on the training they received, and iterate the goals of the program as they understood them as well as how they saw the mentoring program fitting with the district's culture. They also were asked to relate how they work to achieve those goals and what outcomes they are seeing that indicate the program is having a positive effect on new principals and their schools. Similar to the purpose of the administrator and principal protégé interviews, the purpose of these interviews was to formulate a program theory from the mentors' perspectives, compare it to the a priori program theory and the other program theories of stakeholders, and then test its effectiveness based on the leadership growth mentors saw in their protégés over the two years they were partnered.

From these interview data, a culminating program theory was developed from the three participants groups' perspectives (administrators, protégés, mentors). These program theories were compared to each other to explore similarities and differences between groups and analyze whether or not the program administrator's implementation theory is being carried out as intended at the participant level.

Program effects. During phases two and three of this evaluation, I tapped various qualitative and quantitative data sources to identify program effects. First, through the survey and interviews, I asked program administrators, the district superintendent, protégés, and mentors to identify those outcomes that they believed were affected by program participation.

In addition, student data were collected examining student performance since the protégé took over the position. The protégé's length of tenure was taken into consideration when examining these student level data so as to not unfairly give the protégé credit for poor or good performance that may not be due to the protégé's leadership. Student data included measures of student achievement data, including achievement test scores, attendance data, and graduation and dropout rates (as appropriate).

Phase four: Analysis and interpretations of the data. In the preceding description, I outlined the integration of the mixed methods in instrument development and in data collection activities. Further integration of the methods was incorporated in the analysis and interpretation activities. Analysis of the data was ongoing throughout the evaluation study. The survey was analyzed using SPSS to calculate frequencies and means. In addition, open-ended questions were coded for themes, and the prevalence of codes within the data was calculated to determine which themes have greatest representation.

Certain interview questions, teacher surveys, and student data were used to test the a priori program theory and its effectiveness in supporting new principals in their capacity to improve the teaching and learning in their schools. Interview data were coded thematically. Document/artifact reviews and reviews of student performance data consisted of pulling out and coding those pieces of evidence that represent the leadership practices of the new principals.

Criteria for judging program quality. Because this evaluation was conducted within the constructivist paradigm and dominated by qualitative methods, I used the types of validity identified by Maxwell (1996) to examine the trustworthiness of my interpretations. The three types of validity that were most relevant to this study were: descriptive, constructivist, and theoretical. First, descriptive validity refers to the accuracy in the portrayal of the case (i.e., the Clearlake mentoring program) and participants' meanings and perspectives. Second, constructivist validity is concerned with how well I integrated the findings from qualitative and quantitative data sources in ways that accurately portrayed participants' experiences and the effects of the program. Third, theoretical validity questions how well the theory I developed fits the data.

These types of validity account for what is traditionally termed "internal validity" in research and evaluation and determines to what degree an evaluator's warrants can be justified by the data. To control for threats to validity during the interviews, I continuously used member checking to confirm their responses and to check my initial interpretations. During interviews I would mimic back what I thought I was hearing during the interview, as well as describe preliminary interpretations I made based on the responses to my questions. Interviewees corrected my errors and we discussed my preliminary interpretations until we both felt comfortable with my notes. I also initiated triangulation components using multiple methods and multiple stakeholders to confirm and disconfirm findings. Thus, the survey findings were compared to interview data, and interview data were compared among and between participant groups (e.g., administrator responses compared to other administrators', and protégés' responses compared to mentors). Entries in my reflective journal, memos in my notes, and an audit trail of my qualitative and quantitative analysis also provided evidence to substantiate the validity of my

interpretations. Finally, once the report was written, program administrators and participants in the program were given the opportunity to review the report and offer feedback on the interpretations and final program theory. I made corrections to the data or my interpretations where the reviewers identified errors, but mostly I received favorable feedback on the report from the participants who believed that I portrayed the program and participants' experiences accurately.

After this evaluation ended, Illinois Public Act 094-1039 (Illinois General Assembly, 2006) was implemented in FY08 (i.e., beginning in Fall 2007). In accordance with the legislation, school districts in Illinois were required to provide mentoring to new principals in their first year of the principalship by a trained mentor. An advisory committee of educational leadership stakeholders (e.g., Illinois Principals Association, Regional Offices of Education, Illinois Education Association, school district superintendents and principals) was convened to develop the standards and training for the Illinois New Principal Mentoring Program. The Illinois Principals Association (IPA) became the administrator of this program for Illinois. Individuals and organizations who wanted to provide mentoring programs for Illinois new principals had to submit their program's plans to IPA to be approved under the guiding mentoring program standards. School districts, including Clearlake School District, and professional organizations such as the Consortium for Educational Change, and Regional Offices of Education became approved mentoring program providers. Unfortunately, in FY 12 (i.e., Fall 2011), the Illinois State Board of Education did not appropriate funds to the Illinois New Principal Mentoring Program. Thus, the program was no longer mandatory for new principals but was voluntary. Many school districts in the suburbs and Chicago Public Schools continued to provide mentors

for their new principals. However, within time, without state funding, and under new district leadership, Clearlake eventually stopped offering its mentoring program.

Summary of the Findings

This section presents a brief summary of the final evaluation findings. To find the full set of findings and recommendations, the final evaluation report is available upon request. The data collected using the research questions were conveyed in a model of the program's theory. Each part of the program theory was explained and supported by data collected during the evaluation. Figure 7 is a graphical representation of the Clearlake new principal mentoring program theory.

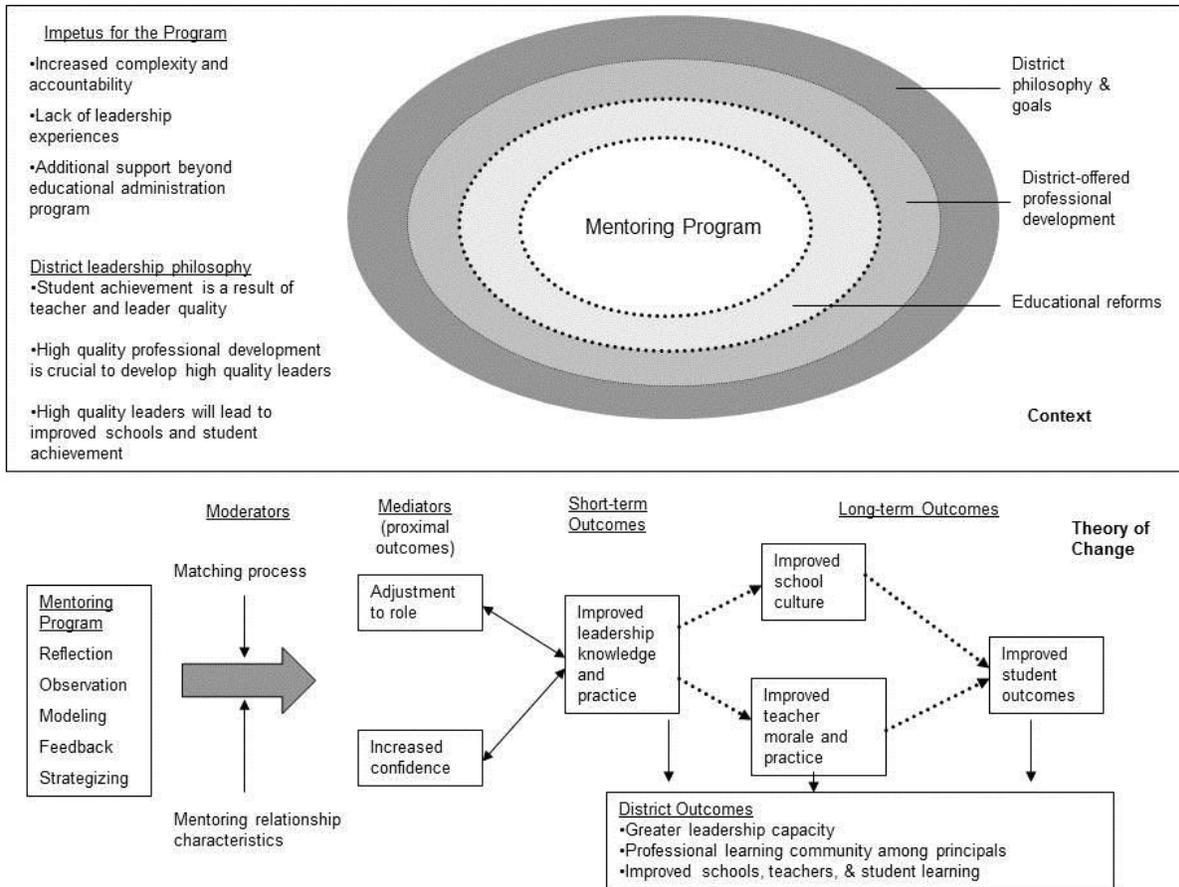
The information contained in the top box set the context of the program which included the philosophy of the district that supports the program, the reasons why the program was developed, and the relation of the mentoring program to other district professional development and educational reform activities.

Under the program's context, the figure depicted the program's theory of change. The first column described the essential activities in which mentors and protégés engaged that led to the program's outcomes. The effects of the mentoring activities were moderated by the effectiveness of the mentor-protégé matching process and the characteristics of the mentoring relationship. For mentors and protégés who developed a successful relationship and engaged in mentoring activities such as reflection and strategic planning, the mentoring program could result in increased confidence and quicker adjustment to the role of the principal. Subsequently, increased confidence and adjustment led to improved leadership knowledge and practices. Improved leadership practices led to improved schools and student learning. For the district, the mentoring program led to improved leadership capacity; a professional learning community of

leaders; and improved schools, teachers, and student learning. Below is a summary description of the program theory of the Clearlake new principal mentoring program.

Figure 7

Clearlake New Principal Mentoring Program Theory



Context

The program theory of the Clearlake Novice Principal Mentoring Program began with a description of the program’s context. First, the mentoring program was situated within a district philosophy that valued school leaders because they were the most important person in the school to foster and implement a shared vision and culture of schooling that was focused on teaching and learning. District administrators chose to implement a mentoring program for several

reasons: a) the new cadre of principals were not coming to the position with much prior leadership experience, b) the district administrators were not satisfied with the quality of preparation programs, and c) the complexity and accountability demands of the principalship requires that novice principals receive additional support while on-the-job.

Finally, the mentoring program was situated within a larger professional development system that the district administrators had in place for administrators and teachers. Some of these professional development offerings were geared toward all administrators, others just for new principals. Other professional development activities were focused on specific educational improvement and reform initiatives in which the district was participating. One of the benefits that new principals reported was that the mentoring program helped make sense and tie together these professional development activities. The mentoring program was especially helped if the mentor and protégé attended the same professional development activities because then the protégés could reflect on what they heard and learned with the mentor and individualize the learning to the particular needs of their schools.

Mentoring Program Process

Beginning in 2002, all new principals in the district were required to participate in the new principal mentoring program. Upon their hiring, new principals were paired with a retired or practicing “master” Clearlake school district principal who served as the new principal’s mentor for their first and second year. A master principal was defined by program administrators as an administrator who has had his/her administrative certificate for at least five years, served as an educator between five to ten years, and had a record of successful leadership. Mentor roles were described as being critical friends who push the protégés beyond their comfort zone. Mentors said their purpose was to supervise, support, and provide technical assistance to new principals.

Mentors and their protégés began their relationship by assessing the principal's strengths and weaknesses against the Interstate School Leaders Licensure Consortium (ISLLC) standards (NPBEA, 2008). From this gap analysis, mentors and protégés set goals for the year. Mentors and protégés were required to meet one hour each week for the protégés' first two years. However, many mentors and protégés indicated that their relationships were more flexible relationships based on the needs of the protégés and what was going on at the school. Mentors were required to keep a log of the meetings to keep track of meeting topics. In addition, some protégés kept their own personal journals to track meeting topics, their reflections on the discussions, and their leadership development. Throughout the year, mentors and protégés alternated their meetings between their respective buildings. This allowed the mentors to observe the protégés' building—the environment (physical and social), the teachers, and students. Another common practice was mentor and protégé classroom walkthroughs. During these walkthroughs, mentors observed teachers in the protégés' schools as they led their classrooms. Afterwards, the mentor and protégé reflected on the teaching and learning they observed.

Program administrators implemented the program with the purpose of developing effective and adaptive instructional leaders for the district's K-12 education system. However, the adaptive instructional leadership sometimes had to follow the managerial role because proper management of the building was still the primary expectation of the faculty and staff. Many of the issues focused on balancing managerial and instructional leadership responsibilities.

Mediators

The model indicated that the first effects associated with this program mentioned most by the administrators, mentors, and protégés were an increase in the protégés' confidence in their leadership abilities and growth in their adjustment to the role of building leader. These outcomes

mediated, or facilitated, the mentoring effects that resulted in the short-term outcomes of improving and learning leadership knowledge and practices.

Mediator: Confidence in leadership abilities. The protégés said that building experience and reflecting on these experiences with their mentors helped build confidence. Knowing that other principals in the district dealt with similar issues helped them feel less isolated and more confident that they could learn from their experiences. Survey respondents said that the mentors helped them feel more confident in areas such as decision-making, problem-solving, building a shared culture, and dealing with staff issues.

Mediator: Adjustment to the position. The mentoring program helped protégés adjust to their role as the building leader. Two of the five case study protégés came straight out of the teacher ranks into the principal position. The other three had spent some time as an assistant principal before becoming a principal. The two former teachers were particularly adamant that the mentoring program helped them make the adjustment from teacher to principal. Survey respondents said that the mentoring program positively affected their adjustment in areas related to being an instructional leader, managing the school, dealing with staff and parents, developing a professional development plan, and implementing district policies.

Short Term Outcomes: Learning New Leadership Knowledge and Practices

Survey respondents said that one of the most beneficial components of the mentoring program was learning new leadership practices during their times of reflection with their mentors. The most common positive impacts on leadership practices were associated with working with teachers, primarily in providing professional development, implementing the School Improvement Plan (SIP), and improving teacher morale. Other leadership practices that protégés said were improved by their mentoring program included: celebrating school

accomplishments, fostering a shared culture, establishing standardized operating procedures, adapting to the leadership needs of their schools, being an advocate for the school, and acquiring knowledge of curriculum and instruction.

Long-Term Outcomes: Effects on Schools, Teachers, and Students

The long-term outcomes that program stakeholders attributed to the mentoring program included outcomes related to improved school climate, improved teaching practices, and finally, improved student achievement. One caveat to this line of outcomes: attributing a link between participation in the mentoring program to these more distant outcomes is tenuous and based on the beliefs of the administrators, mentors and protégés. Data collection efforts attempted to corroborate their beliefs with data from teachers, but the response rate to teacher surveys was too low to be able to make any conclusions.

Administrators and protégés claimed that new principals were being more thoughtful and transparent in the ways that they worked with teachers. Protégés said that in their schools they created climates that were transparent in their expectations and processes and transparent in presenting data to faculty and staff. Protégés were working with teachers to build morale and building a shared vision of schooling with teachers. As a result of this work, protégés reported higher teacher morale. Mostly they have seen teachers return to a vision of why they began teaching in the first place, to impact students' lives. Protégés also reported seeing teachers set higher expectations for themselves and put student learning at the forefront of their work. Finally, study participants were asked what they think the effects of the mentoring program have been on students in the protégés schools. Administrators claimed that in 2006 (the most current assessment test data at the time of this evaluation), all of the district's elementary schools made Adequate Yearly Progress (AYP), while the middle schools missed AYP targets in just one

subgroup. Protégés felt that the effects on students were trending in a positive direction. Because the protégés have begun using data to gauge student learning, they are quicker to see how students are progressing in different areas. An examination of the protégés' school report cards for the years before, during, and after (when applicable) participating in the mentoring program revealed that the scores were variable; there was not a consistent upward trajectory in student achievement. In the interviews with the protégés, they revealed that they were dealing with difficult issues, particularly in the lowest income and highest minority schools, that will need more time and work to overcome in order to raise student achievement.

Effects on the District

Program stakeholders reported that the mentoring program is helping build the district's leadership capacity by helping them become better principals and more confident in their skills. In addition, administrators said that the mentoring program has helped created a larger learning community among the district's principals. Protégés echoed the administrators' reflections and said that now knowledge is being shared across the district. It has become the culture of the district that it is okay to share knowledge and support each other. Finally, program administrators said that new principals are engaging more in instructional conversations rather than sitting on the sidelines and leaving these discussions to teachers. Administrators reported not seeing the typical crises that most principals face during their first year (e.g., poor school management; teacher, staff, and parent issues) for which new principals often seek help from the district. Thus, district administrators reported that they received fewer complaints against the principals (protégés) through phone calls and office visits than they had received about past principals in the same stage of the principalship.

Moderators

There were two sets of moderators listed that influence the size of the effect that the mentoring program had on the protégés. Moderators are those factors internal or external to the program that can affect the size of program effects. The first set of moderators is related to the characteristics of the mentors and protégés and the second set are those factors related to the matching of the mentors and protégés.

One of the most critical factors for protégés was that the mentor did not serve in an evaluative role and all discussions between mentor and protégé were held confidential. Therefore, protégés believed that they could expose their weaknesses without worrying about affecting their annual performance evaluations. They believed it was most beneficial to have mentors with whom they could reflect on their leadership practice and learn about resources where they could find helpful solutions.

At the beginning the mentoring program was a program without a formalized model and there was very little to no training, no expectations for how often mentors and protégés should meet, and little guidance for conducting the role of mentor. In 2004, the program adapted the Coaching Leaders to Attain Student Success (CLASS) coaching program developed by the New Teacher Center at the University of California at Santa Cruz. With this program, the training became much more structured using a three-day format in which mentors were taught the CLASS model by a program coordinator at the Consortium for Educational Change (CEC). Mentors said that providing mentor training based on a model was critical as the district recruited new mentors. Mentor training taught them what good mentoring looked like and gave mentors a common set of mentoring practices that guided their work.

In terms of the mentoring relationship, mentors and protégés reported that it was important that the relationship was characterized by commitment, openness, trust, and respect. Both groups said that the protégés and mentors must be open individuals and unafraid to ask questions. They also said that protégés must be willing to expose their vulnerabilities and weaknesses, and mentors must be willing to share their ideas and leadership strategies with their protégés. It was crucial that protégés feel that they could trust that their mentors would not share their protégés' weaknesses, or gaps in leadership development, with others in the district, including the district administrators who supervised and evaluated the principals in the district.

The second set of moderators related to the characteristics by which mentors and protégés were matched. This set included matching by school grade level, mentor and protégé strengths and weaknesses, school demographics, and school/community issues. Protégés said that it was important to match mentors and protégés by school grade level because they were not convinced that a mentor from a high school could help a protégé from an elementary school (or vice versa). There was also an issue of matching mentors and protégés by student demographics because the challenges of principals from schools with different student demographics could be quite dissimilar. Mentors also said that was important that the personality between the protégés and mentors match, as well as the strengths against the weaknesses.

Recommendations

The evaluation report ended with recommendations that would help the Clearlake mentoring program administrators improve their program. These included:

- Program administrators should develop a set of well-defined indicators and measures of effectiveness to evaluate their program and make improvements as needed.

- The program administrators should document criteria for screening, selecting, and matching mentors with protégés so that future administrators will have formalized mentor recruitment and matching process in place.
- Program administrators should offer a refresher training course for mentors during the school year.
- Program administrators should consider offering meetings for the mentors and protégés so that they can learn from fellow mentor and protégés about the strategies that are working in other schools.

Metaevaluation of the New Principal Mentoring Program Evaluation

This evaluation served three purposes: 1) to construct knowledge on the effects of mentoring programs on novice principal's leadership development; 2) to provide information to Clearlake administrators for their use in making programmatic changes; and 3) to serve as the case study, or metaevaluation, of a constructivist program theory evaluation approach. Therefore, there were two overlapping studies: the evaluation study, and the metaevaluation. This dissertation primarily is concerned with questions related to the reflective metaevaluation which is further described in chapter four.

Chapter 4

Metaevaluation Study Design

This dissertation is a reflective metaevaluation of the evaluation of the Clearlake School District new principal mentoring program that was conducted using the program theory evaluation model within the constructivist paradigm. My decision to conduct a constructivist program theory evaluation (C-PTE) was influenced by the critics of PTE, particularly Dahler-Larson's (2001) call to integrate constructivist evaluation with program theory evaluation to bring forth the interdependent nature of the program and context and both understand how the program affects stakeholders and the environment and vice versa. My critique of PTE also parallels the work of Charmaz (2000, 2005, 2006) to develop a constructivist grounded theory approach based on her critiques of Glaser and Strauss' (1967) grounded theory methods.

Recognizing that PTE and grounded theory methods are not the same approaches to evaluation and research, respectively, the two do share a similar focus on theory and philosophical traditions. PTE is a theory of how to conduct an evaluation that is driven by the evaluand's program theory, but is not a prescribed set of methods to develop and test the program theory. On the other hand, Glaser and Strauss (1967), and later Strauss and Corbin (1990), developed grounded theory methods as a prescriptive set of methods to collect and analyze data and develop a theory grounded in the data. PTE and grounded theory methods do share a focus on developing a theory, grounded in data on specific social phenomena. In addition, both of these approaches to research/evaluation have their roots in post-positivist philosophies of ontology, epistemology, and representation of the program. In addition, evaluators who operate within the PTE and grounded theory approaches regard programs as external phenomena to be discovered by the objective inquirer who remains external to the

program. The methods the researcher/evaluator use allow the evaluator to remain objective and unbiased. Grounded theorists use methods of constant comparative analysis to remain objective (Glaser & Strauss, 1967), while many PTE evaluators have used quantitative statistical modeling techniques to isolate, define, and measure variables to provide a conceptual, explanatory link to explain their relationships to one another (Bickman, 1996; Reynolds, 1998; Trochim & Cook, 1992).

To begin, postpositivism and constructivism differ in their accounts of reality. Charmaz (2005), whose work differs from Glaser, Strauss and Corbin in that her work is explicitly constructivist, said that in constructivist grounded theory inquirers bring their histories, prior experiences, and philosophical stances with them into the inquiry event. While postpositivists claim that there is an external reality that is observable, independent of the inquirer, and guided by natural laws, constructivists in contrast, believe that there are multiple socially constructed realities, dependent on the interactions between the observer and the observed, ungoverned by natural laws (Guba & Lincoln, 1989). Thus, constructivism claims that inquirers are not impartial observers but that our background and philosophies frame the questions we ask and serve as a lens through which we observe the world. From our inquiry we develop representations of the data (e.g., program theory) that are constructivist depictions of the data rather than objective accounts of reality. Finally, Charmaz (2005) suggested that researchers adopt a reflective stance in their research activities, acknowledging the influence their prior knowledge and experience has on the questions they ask and the interpretations they make from the data.

With the constructivist paradigm in mind, during the evaluation I worked under the following assumptions. First, data in this evaluation were a collection of lived experiences and values and the “realities” of the program were the constructions of the program participants,

including the program administrators, principal protégés, and mentors (Guba, 1987; Guba & Lincoln, 1989). Thus, in lieu of the search for “facts,” this evaluation was concerned with eliciting program participants’ experiences, meanings, and values associated with the mentoring program (Geertz, 1973; Guba & Lincoln, 1989). The data collection activities (e.g., interviews, surveys, and document reviews) were methods of drawing out the participants’ experiences and values associated with the program. As the evaluator, my primary role was to facilitate these constructions and act as a co-constructor of these experiences with the viewpoint that no constructions were “right or wrong” but rather represented subjective recollections of those aspects and experiences of the program that the participants deemed most salient or important (Guba & Lincoln, 1989). The participants’ responses to my questions were used to construct program theories, and therefore, I anticipated the construction of multiple program theories that would reflect multiple constructions and values associated with the mentoring program. I also worked under the assumption that programs evolve and adapt to the needs of the environment (Dahler-Larson, 2001; Guba & Lincoln, 1989). Thus, I assumed that the mentoring program as implemented today may not be the same program that was implemented at its inception in response to the changing needs of the district. Documenting this evolution of the district mentoring program was a priority in this evaluation. The benefits and challenges associated with blending constructivist principles with the program theory evaluation model was a focus of this dissertation study. In essence, I was interested in whether integrating these two types of evaluation theories would help me create an evaluation that brought in the richness and complexity of programs and program experiences and be useful in helping the reader learn more about how mentoring programs operate and whether they are a valued professional development strategy for school districts.

Research Questions

This dissertation is a metaevaluation of a constructivist program theory evaluation of a district new principal mentoring program conducted within a primarily a constructivist qualitative design. Some of the concepts and assumptions of program theory evaluation and constructivist evaluation are inherently incompatible. Often, program theory evaluation is learning about programs at the macro-level, developing information that is useful for policy decisions. Some evaluators identify program theory a priori to guide the evaluation, using the theory to frame the questions, choose methods, and analyze and interpret the data (Chen, 1990; Greene, 1993). Constructivist evaluation, in contrast, is emergent and responsive to the needs and issues that arise in the program context, has a utilization focus often primarily aimed at local stakeholders, and concerns itself with local program relevance versus relevance at a macro-policy level (Greene, 1993). Thus, this study is a critical reflection of the issues and tensions and benefits related to blending program theory evaluation into a qualitative, constructivist evaluation design guided by the following questions:

1. How did integrating elements of constructivist and program theory evaluation affect the quality of the evaluation in terms of:
 - a. Constructing knowledge about how the district mentoring program works;
 - b. Representing the multiple theories of stakeholders;
 - c. Describing the relationship between the program and its context; and
 - d. Identifying the program's effects on principals, schools, and students?
2. What is the utility of a constructivist program theory evaluation for learning about principal mentoring programs in terms of:

- a. Helping program administrators understand mentoring programs and make changes or improvements to their own programs;
 - b. Informing the leadership field about how mentoring programs work and the effects they have on principals and their schools?
3. How does this evaluation compare to a sample of constructivist, program theory, and realist evaluations in quality and utility? Did those evaluations meet the quality and utility criteria to a greater or lesser extent than mine, or did they face many of the same or additional challenges?

The remainder of this chapter outlines the design of this study to explore these research questions.

Study Design

As a metaevaluation, this dissertation was primarily concerned about the quality of the evaluation of the mentoring program as well as the usefulness of the information for learning about principal mentoring programs. To answer the research questions, I used three methods: 1) reflective analysis, 2) comparative analysis of the C-PTE to exemplars of PTE and constructivist evaluations, and 3) a metaevaluation panel review, which consisted of evaluators and mentoring program administrators.

Quality Criteria

To reflect on and evaluate the quality of this evaluation in terms of the inferences, I used a mix of constructivist quality criteria as defined by Guba and Lincoln (1989) in *Fourth Generation Evaluation*, criteria standards from realist evaluation (Julnes & Mark; 1998; Mark, Henry, & Julnes, 1998), and utility criteria from the Joint Committee on the Standards for

Educational Evaluation, 2nd edition (1994)⁵. While not a specific focus of this metaevaluation, the development of these criteria is itself an important contribution of this dissertation and will be discussed further in the last chapter. These criteria framed my analyses of my reflective journal entries as well as the metaevaluation panel responses and comparative analysis of this evaluation with exemplar evaluations and research, including PTE and constructivist evaluations, and research and evaluation in the school leadership field. Furthermore, these criteria were used by the metaevaluation panel to metaevaluate the constructivist PTE.

In order to use these criteria for this metaevaluation, I defined and operationalized each criterion so as to standardize the analyses across methods. For example, I defined and operationalized the criterion of credibility so that my definitions and coding on credibility issues are similar in my reflective analysis and the analysis of the external panel review. After operationalizing each criterion, an interview question, or set of interview questions, was developed to elicit feedback from the experts on the panel review on how my evaluation report measured up to it.

Question one focused on evaluating the constructivist PTE on five dimensions of quality. To address the constructivist components of the evaluation, I employed constructivist criteria of trustworthiness and authenticity (Guba & Lincoln, 1989). The trustworthiness criteria are “parallel criteria” to conventional criteria that examine the rigor of evaluations and the authenticity criteria are focused on the extent to which an evaluation represented the constructions of stakeholders (Guba & Lincoln, 1989). The purpose of these criteria are to metaevaluate the quality of the constructivist PTE in terms of the quality of inferences, its

⁵ A 3rd edition of The Joint Committee on the Standards for Educational Evaluation was published in 2010. I used criteria from the 2nd edition and kept the citation from this edition to retain the integrity between the criteria used and the original source.

explanation of the relationship between the program and its context, and the quality with which the evaluation represented the program theory from multiple perspectives of the stakeholders.

The trustworthiness and authenticity criteria that I used to evaluate this constructivist PTE included: credibility and fairness. Credibility parallels internal validity but credibility assesses the extent to which the evaluator's program constructions correspond to the constructed representations of stakeholders (Guba & Lincoln, 1989). Fairness is the extent to which the evaluator fairly represented the program theory of multiple stakeholders in terms of the reporting and prioritizing of different viewpoints, and the negotiation of multiple constructions into the final program theory representation (Guba & Lincoln, 1989). To complement these criteria, I added a quality criterion that I named context-program interactions influenced by Dahler-Larson's (2001) writing on blending constructivism and program theory evaluation. This criterion addressed the issues of the interdependency, or conditional relationship, of programs and the contexts within which they reside. As such, the metaevaluation panel and I examined the extent to which the evaluation I conducted of the mentoring program attended to the effects of the context on the mentoring program and conversely, the effects of the program on its context.

Furthermore, I brought in concepts of realist theory to examine the quality of inferences related to the causal effects attributed to the program. In chapter one, I gave multiple definitions of program theory in which program theory is the underlying beliefs and assumptions that describe how and why a program works. More specifically, PTE identifies the most effective program components as well as the program's mediating causal mechanisms by which the program achieves its effects, and the moderating mechanisms that affect the degree of program effects (Donaldson & Lipsey, 2006). Complementary to PTE, realist evaluation has as its focus:

identifying generative mechanisms that underlie causes, program conditions under which the conditions operate, and for whom the mechanisms operate (Mark, Henry, & Julnes, 1998).

In addition to the compatibility between realist theories with PTE, realist philosophies also are compatible with constructivist philosophies of how humans construct reality. Realist philosophical assumptions encompass constructivist beliefs that reality is a construction that is subjectively experienced by humans and that our methods for observing reality are constructed (Mark et al., 1998). Both theories posit that there is a social world that exists, but that we can only know it from our own standpoint, thus our knowledge of the real world is constructed. Realist theory states that program effects are caused by both observable processes and unobservable underlying mechanisms and dependent on contextual features that impact the effects' intensity. Therefore, I chose to incorporate realist criteria as a fairer judge of the quality of causal inferences from the constructivist PTE than a metaevaluation that is performed using only constructivist criteria.

Realist theory states that evaluation has two foci: sensemaking and valuing (Julnes & Mark, 1998; Mark et al., 1998). Constructivist philosophy also focuses on valuing in relation to the data and findings of evaluation, and the constructivist criteria that I chose for this evaluation covered the valuing aspect. Therefore, for this metaevaluation, I brought in criteria related to sensemaking. Sensemaking, according to Mark et al. (1998) is an evaluator's attempt to construct knowledge about a program or practice, to construct order out of the patterns that we observe. These realist evaluation theorists stated that the difference between realist theories notion of knowledge construction versus the post-positivists is that realist evaluators assume that humans have a natural ability to make sense of the world, while positivists rely on formal methods of observation.

The realist criteria that I included in this metaevaluation are: competitive elaboration and principled discovery (Julnes & Mark, 1998). Competitive elaboration describes the extent to which the evaluator identified alternative explanations to attribute program effects and collected data to rule these alternatives out. This criterion suggests that prior to or early in the evaluation, evaluators should identify possible alternative explanations and devise a plan to study these alternatives. Failure to do so may result in erroneous inferences in relation to program effects. Principled discovery is the extent to which an evaluation builds theory in a systematic manner to explain a finding that might have otherwise been overlooked. Principled discovery relies on inductive and deductive reasoning to search for confirming and disconfirming explanations of program effects. Thus, adding realist criteria to the metaevaluation of the constructivist PTE allowed me to explore the contributions of the program theory in identifying causal linkages in an otherwise constructivist, qualitative evaluation. It also allowed me to explore the contributions of constructivist, qualitative evaluation to observing causal patterns and making valid causal inferences.

Utility Criteria

Question two is concerned with the usefulness of the evaluation findings for learning about and managing principal mentoring programs. To metaevaluate the constructivist PTE on its utility, I blended the utility criteria from the Joint Committee on the Standards for Educational Evaluation (1994) with elements of Guba and Lincoln's (1989) constructivist evaluation, and one criterion adopted from Charmaz' writing on constructivist grounded theory. In terms of evaluation utility, I used the criteria of information scope and selection (Joint Committee on the Standards for Educational Evaluation, 1994), and values identification (Joint Committee on the Standards for Educational Evaluation, 1994), report clarity (Joint Committee on the Standards

for Educational Evaluation, 1994), transferability (Guba & Lincoln, 1989), and resonance (Charmaz, 2006). Information scope and selection is defined as the extent to which the evaluation collected information that was broadly selected to meet the information needs of a broad set of stakeholders. Values identification judged the evaluation on how well it clearly described the stakeholder perspectives, procedures, and rationale that were used to draw inferences from the data. Furthermore, the evaluation report was evaluated according to the clarity of its presentation of the program, its context, and the evaluation findings. I included this criterion to rule out the possibility that the report was unclearly written and influenced the metaevaluation panel's responses on the questions of quality and utility.

On the constructivist side, I incorporated one criterion: transferability (Guba & Lincoln, 1989). Transferability is the extent to which the evaluation findings can be generalized to another program and context. The judgment on this criterion is made by an external audience who determine whether or not the evaluation findings reflect their circumstances. Thus, this was a criterion judged by the interviews with program administrators in which I asked them to what extent the evaluation findings match their own understanding of their mentoring programs.

To further complement constructivist criteria, I included Charmaz's (2006) criterion of utility she called resonance. In particular, I explored the extent to which the program theory made sense to and presented new and deeper understanding of principal mentoring programs for program participants and administrators, as well as to the mentoring program administrators who served on the metaevaluation panel.

Methods

This study was guided by three research questions that asked about the quality and utility of the constructive program theory evaluation of a new principal mentoring program, and how

this evaluation compared to other similar evaluations in terms of quality and utility. Table 2 lists the methods used to explore these questions, their purpose, the quality criteria, and modes of analysis. The remainder of this section provides fuller description of these methods.

Table 2

Study Methods Table

Method	Criteria	Analysis	Purpose
Question 1: What is the quality of this C-PTE?			
Reflective analysis	Credibility Context-program interactions Fairness	Thematic coding	To metaevaluate inference quality, contextual grounding of the program, and representation of the program theory
Metaevaluation panel	Principled discovery Competitive elaboration		
Question 2: What is the utility of this constructivist PTE for learning about principal mentoring programs?			
Reflective analysis	Resonance Transferability Information scope and selection Values identification Report clarity	Thematic coding	
Metaevaluation panel		Thematic coding	To metaevaluate the usefulness of the evaluation for learning about and managing principal mentoring programs
Evaluators	Resonance Values identification Report clarity		
Mentoring Program Administrators	Resonance Transferability Information scope and selection Values identification Report clarity		(Continued)

Table 2

Study Methods Table (Continued)

3. How does this evaluation compare to a sample of evaluations on criteria of quality and utility?			
Comparative analysis	Credibility Context-program interactions Fairness Principled discovery Competitive elaboration Resonance Values identification Report clarity	Thematic coding	To assess whether similar program evaluations were of better, worse or similar levels of quality and utility as this C-PTE

These questions sought to explore the issues and tensions related to blending program theory evaluation with constructivist concepts that value multiple meanings, beliefs, and values of the stakeholders and the emergent and evolving nature of programs and evaluations of these programs. To explore the research questions, I used three methods: a reflective, critical analysis of my experiences with constructivist program theory evaluation using data from my reflective journal, a comparative analysis of my evaluation results against exemplars of program theory evaluation and qualitative evaluation, and a metaevaluation by a panel of reviewers that consisted of two groups of experts—experts in program evaluation and experts in administering new principal mentoring programs.

Method 1: Reflective analysis. During all stages of the evaluation, I maintained a log and reflective journal of the process. This log is a compilation of my notes and reflections on the evaluation process, particularly in regards to the effects of incorporating PTE into the design of the study. The purpose of this analysis is to reflect on the challenges and benefits I encountered when integrating constructivist evaluation and PTE concepts to construct credible and

informative evaluation findings. The purpose of my journal entries was to reflect on my process of designing and implementing a constructivist PTE that resulted in high quality evaluation data. This analysis reflected my work on the evaluation design, instrument development, data collection, analysis, and interpretations (including representing program theory). These reflections included my field notes and memos that revealed my interactions with stakeholders and the data, points of decision, and reflections on the evaluation process as it unfolded. To guide my analysis, I developed a self-reflective analysis protocol guide included in Appendix E. This protocol guide operationalized the quality criteria through a set of questions similar to the ones included in the protocols used during the interviews with the evaluators and mentoring program administrators who participated in the metaevaluation panel. An outside reviewer was recruited to validate my interpretations of my reflective analysis. This person received training and is experienced in qualitative data analytic techniques. I gave this person a sample of my journal entries along with the self-reflective analysis protocol guide that I used to analyze my journal entries. This person coded and verified the findings of my analysis. To conduct this verification process, I gave my colleague a sheet that contained my codes and a description of those codes. My colleague used this document to conduct a coding of the entries, and then the two of us compared the results of her coding with mine. Where there were discrepancies we discussed the differences, and when deemed appropriate, I changed my codes in my reflection journal.

Advocating for reflective writing as a research practice, Jasper (2005) said that the purpose of reflective writing is for learning that triggers a change in action. Furthermore, it provides a decision-trail that highlights the decisions that were made in the conduct of the research. Reflective writing allows the researcher to make new connections of meaning. Finally,

Jasper identified a secondary role whereby reflective data are used as a secondary data source as reflective writing provides the commentary and interactions and the interpretations of the data as the research progressed. Richardson (2000) stated: “Writing as a method of inquiry, then, provides a research practice through which we can investigate how we construct the world, ourselves, and others...” (p. 924). I kept the journal to reflect on my own philosophical stances; substantive knowledge of principal preparation and professional development, education, education reforms, school leadership, and mentoring programs; as well as my personal biases toward the school district and district personnel. In addition, Richardson and St. Pierre (2005) approach writing as method of data collection and a method of data analysis. For example, as a method of data collection, my journal entries consist of my feelings of confusion about how to develop a question that is open-ended, but yet will elicit the participants’ rendering of the mentoring program’s theory, and of my reflections on a frustrating interview in which I wondered if I had successfully captured the information I needed. Through my writing, I began analyzing the data I was hearing during the interviews and forming preliminary program theories in my journal. In these entries, I grappled with where all of the pieces of the mentoring program fit into a theory.

Method 2: Metaevaluation panel review. Finally, for the purpose of addressing the quality of the district mentoring evaluation, I recruited an expert panel of program evaluators and mentoring program administrators to metaevaluate the evaluation of the district mentoring program. Stufflebeam (2001b) defined metaevaluation as the “delineating, obtaining, and applying descriptive information and judgmental information about the utility, feasibility, propriety, and accuracy of an evaluation...to guide the evaluation and/or report its strengths and weaknesses” (p. 185). For this metaevaluation, the metaevaluation panel membership consisted

of four experts in evaluation and three administrators of principal mentoring programs. The evaluators on the panel were chosen because of their history and publishable record of experiences in the evaluation field. The panel included evaluation theorists and practitioners who are adherents to program theory evaluation and constructivist evaluation, as well as evaluation theorists and practitioners who have written and/or presented thoughtful reflections on evaluation theory and practice. The panel was identified by a review of their publications in peer-reviewed journals and their evaluation reports.

The mentoring program administrators were chosen from a list of approved State of Illinois principal mentoring program administrators. The administrators who participated in this panel operate their programs out of a teacher and principal professional development organization, a regional office of superintendents, and an intermediate service center in the Chicago area. I chose this sample because of their deep knowledge of mentoring programs and their years of experience in developing, implementing, and refining these programs. I was particularly interested in whether or not they valued learning the mentoring program's theory and representation of it as a tool for learning about mentoring and for revising their programs. The interview protocol was a series of closed-ended and open-ended items in which program administrators were asked to rate the utility (i.e., information scope and selection, values identification, and report clarity) and transferability of the findings as well as discuss how the findings might be used to inform program decisions. First, I piloted the protocol with a small sample of education evaluators/researchers and mentoring program administrators, professional development providers, and education administration faculty prior to its use with the final sample. I made refinements to the interview protocol based on the pilot sample's feedback.

The panel was given the final evaluation report of the mentoring program evaluation. No identifying information of the district or the program stakeholders who participated in the evaluation was included in this report to protect their anonymity. The panel also received the metaevaluation interview protocol instrument that contained the criteria on which they judged the quality and utility of the evaluation. I constructed a rating and interview instrument using the quality and utility criteria to metaevaluate my constructivist PTE. I developed a rating instrument that corresponds to the criteria and the panel was asked to respond to what extent the evaluator met a standard of the quality and utility criteria (e.g., “to no extent, to some extent, or to great extent”) and provide a rationale for their rating through qualitative comments to a set of questions that operationalized the criteria. The interview guides used during the metaevaluation review are included in Appendices F (interview guide for program evaluators) and G (interview guide for mentoring program administrators). The protocols for the two types of groups in the metaevaluation panel differed slightly. The program evaluation experts on the panel answered to the quality and utility criteria of: credibility, context-program interactions, fairness, principled discovery, competitive elaboration, resonance, values identification, and report clarity. The mentoring program administrators on the panel answered questions for these same criteria but also two extra utility criteria to which the evaluators did not respond. These were: transferability and information scope and selection. To be able to discuss my evaluation’s meeting of these two criteria necessitated a deeper knowledge of the educational leadership development field in general, and principal mentoring programs in specific. This was knowledge that the program evaluators on my panel may not have had the background knowledge to speak to these two criteria in the depth that I desired for the metaevaluation..

Method 3: Comparative analysis. To add to my critical reflection on this evaluation, I chose five examples of evaluations that were conducted using the following approaches to evaluation: PTE, realist evaluation, interpretive PTE, and constructivism. A brief description of these evaluations is included in Appendix H. The purpose of this analysis was to compare these evaluations on the same quality and utility criteria used by the metaevaluation panel and in my reflective analysis to critique my evaluation of the new principal mentoring program. To identify evaluations for this comparative analysis, I asked the metaevaluation panelists who reviewed the district mentoring program evaluation to recommend potential exemplar PTE and constructivist evaluations. The panel nominated exemplar program theory evaluations and constructivist evaluations that they believe best represent PTE and constructivist evaluation. I also searched databases such as ERIC and Dissertation Theses and Abstracts, as well as the journals *American Journal of Evaluation* and *New Directions in Evaluation* using the keywords: program theory evaluation, constructivism, interpretivism, and realist/realism evaluation. The final selection of evaluations were chosen based on the transparency of the process wherein I could clearly see the evaluation design, methods, data analysis techniques, and particularly the evaluators' representation of the findings (e.g., the identified program theory or narrative).

In conducting the comparative analysis, I constructed a protocol that was an adaptation of the interview protocol used to guide the discussion with the metaevaluation panel. The comparative analysis protocol is included in Appendix I. For this analysis, I compared a sample of program evaluations to mine using the quality criteria and utility criteria identified in Table 2. To conduct the comparative analysis, I used the protocol used to interview the metaevaluation panel so that I essentially engaged in the same task that I asked of my panel. For each program evaluation report, I read the report and answered the questions on the protocol. Like the

metaevaluation panel's questions, the questions began with "to what extent" type questions for each of the criteria and included open-ended questions that provided space to elaborate on how well I believed the program evaluations met the criteria.

Data Analysis

In respect to analyzing the content of my reflective journal, the material was organized according to the phases of evaluation study the entry addressed. There is a large amount of data contained in these journal entries, and the first task was to organize the data to facilitate analysis. Thus, the entries were organized according to the following phases: evaluation design, instrument development, data collection, analysis, and interpretations (including representing program theory). After the entries were organized categorically, journal entries were coded and analyzed thematically using qualitative coding and categorization methods (Coffey & Atkinson, 1996) and guided by a self-reflection protocol based on the quality and utility criteria described above. The issues that were salient in my mind as I coded the data were those issues related to the quality and utility of the evaluation, and the tensions and issues related to blending PTE and constructivism. To begin, the data were reviewed as I took notes on preliminary codes and categories related to the criteria that emerged from the journal entries. After this initial scan of the data, I reviewed my notes and developed a coding scheme to guide further analysis. The analysis began at a descriptive level, identifying the common characteristics that described the data within the codes, but moved to a more relational, holistic level as I connected themes across journal entries to provide critical interpretations of the issues and challenges I faced in conducting the evaluation. An external reviewer was given a copy of my journal entries and asked to conduct an audit trail to validate my interpretations. The data collected through the metaevaluation panel review were analyzed using the same set of coding and categorization

strategies using the quality and utility criteria to create and organize the codes and resulting themes.

To analyze the results of the metaevaluation panel, I conducted two analyses. The first qualitative analysis was a comprehensive analysis of responses to the protocol questions across all panelists. The second analysis juxtaposed the responses by panelist group (i.e., evaluators' responses versus mentoring program administrators' responses). The responses were coded for each of the quality and utility criteria and themes were developed from groupings of the codes.

Finally, for the comparative analysis, the results of my review of the exemplar evaluation reports were compared across the evaluations and codes were developed for each of the quality and utility criteria. In this analysis, codes were developed and used to formulate themes. These were then compared to the results of my reflective analysis and the metaevaluation panel analysis. This analysis identified common and disparate themes across the three data sources (i.e., the reflective analysis, the metaevaluation panel analysis, and the comparative analysis).

Drawing Conclusions and Verifying the Data

During the data collection and analyses phases of this study, many efforts were made to assure the trustworthiness of my conclusions. First, to establish warrant for my conclusions from the reflective and comparative analysis, I strove for transparency in my methods and data sources, providing ample quotations and examples from my journal entries to substantiate my claims and to allow the reader to either agree or disagree with my conclusions. In addition, I recruited an outside person to analyze my journal entries to warrant my interpretations by conducting a confirmability audit. I gave this auditor samples of my journal entries to confirm that my interpretations were rooted in the data collected in the journal. Furthermore, to establish credibility in regards to the conclusions drawn from the metaevaluation panel metaevaluation

and the survey to mentoring program administrators, I engaged in two techniques suggested by Guba and Lincoln (1985): member checking and triangulation. In order to warrant the conclusions drawn from the responses in the panel metaevaluation and administrator interview, I shared my preliminary interpretations of the panel members' responses to determine if my conclusions reflected their perceptions of the evaluation. The panel members corrected me where my interpretations fell short. Mostly, the panel members confirmed that my interpretations were in line with their statements during our interview.

Creswell (1998) described triangulation as the "searching for the convergence of information" (p. 213). Most commonly, different sources of information are used in qualitative research as a strategy to verify the information (Lincoln & Guba, 1985). In this instance, the overarching thesis of this dissertation is whether integrating program theory evaluation into a constructivist qualitative evaluation design produced a quality and useful evaluation. Therefore, in this case, all four methods of this study triangulated on this overarching question.

Transferability. In constructivist evaluation, the purpose is often not to generalize results, but instead to contextualize information, facilitating a deeper understanding of a phenomenon (Kvale, 1996). Qualitative evaluators often provide thick descriptions that allow the reader to make the judgment of the transferability of the data to other case examples (Geertz, 1973). In the case of this study, I provided thick description of my reflections and analyses of the evaluation and the tensions with which I wrestled as I blended two paradigms, not as a means of critiquing PTE on the whole, but to provide thick description of my practice as an evaluator as I grappled with two sometimes competing theories of evaluation, reporting both the benefits and the challenges. As such, because this is one study of one evaluation using PTE, this case study is limited in its generalizability as a critique of PTE. This case study provided depth through my

reflections of the case, the review by the metaevaluation panel, and the comparative analysis. However, this evaluation was my own and my reflection and the metaevaluation panel's reflections on the evaluation process and results were specific to this evaluation process. The purpose of this dissertation study was not to generalize to all cases of program theory evaluation, but to add to the dialogue and empirical knowledge of evaluation practice.

Evaluator Stance

On the whole, I tend to be more interested in those questions best explored through a constructivist lens. These are typically questions that value depth over breadth. In addition, I value inclusivity in my evaluation practice in which the voices and meanings of multiple stakeholders are sought out with the intention of bringing a diversity of beliefs, values, and experiences into the evaluation. Furthermore, I am an advocate for constructivist, qualitative methodologies and look for ways of making qualitative evaluations more relevant and valued by educators and policy-makers. In respect to the district mentoring program and its relationship to the educational leadership field, I see a strong role for qualitative methods in disentangling claims of program impact, especially in the field of educational leadership in which impacts on schools and student achievement are often not direct, but where immediate outcomes on principal development and practice can lead to a transformation in school culture and practices, which lead to higher quality instruction, and subsequently to the ultimate impact of improved student achievement.

Summary

In this chapter, I presented the research design for this study that explores the issues and challenges of conducting a program theory evaluation within the constructivist paradigm. This study was a metaevaluation of a constructivist program theory evaluation of a district's new

principal mentoring program. The criteria used to judge the evaluation came from constructivist criteria (Guba & Lincoln, 1989), realist theory (Julnes & Mark, 1998; Mark et al., 1998), and the Joint Committee on Standards for Evaluation (1994), as well as the writings of evaluation practitioners Charmaz (2006) and Dahler-Larsen (2001). The study consisted of a reflective and comparative analysis and a metaevaluation panel review of the evaluation that examined the issues related to the quality of the findings and inferences, the extent to which the program was linked to its context, and the accuracy of its representation of the program's theory, as well as to discuss the usefulness of the constructivist PTE in providing information to help mentoring program administrators manage and improve their mentoring programs.

Chapter 5

Results of the Metaevaluation

This chapter presents the findings of the data analysis that reflected on the quality and utility of my evaluation of the Clearlake School District's new principal mentoring program. The following sections capture the dialogue among the three sources of data used to metaevaluate the evaluation: the analysis of the reflection journal entries, metaevaluation panel responses, and the comparative analysis.

Quality of a Constructivist Program Theory Evaluation

The following section highlights the findings of the analysis of the reflective journal and the metaevaluation panel interviews. For the quality criteria, the data in my journal and the interviews are in response to questions about credibility, fairness, context-program interdependence, principled discovery, and competitive elaboration.

Credibility: Presenting strong linkages between the data and the evaluation's claims. Upon analysis of the reflection journal entries written when I designed the evaluation, one of my primary concerns was designing an evaluation that honored the complexity of mentoring programs specifically as well as the complexity of schooling. In other words, I wanted to capture the contexts within which schools and learning reside as well as how educators improve their knowledge and practice as well as improve schools and student learning. In the first journal entry, I reflected on why I chose program theory evaluation to address the need for identifying a causal chain of effects. I wrote:

Researchers for over 10 to 15 years have attempted to correlate principals' practice and their effects on school culture, teaching, and student learning. Researchers have found a small but significant link that principals can affect student learning either positively or

negatively. Taking this one step further, researchers and practitioners would like to see this chain of effects to incorporate how principal preparation and professional development programs affect principals' practice and then the impact on schools, teachers, and students.

A couple paragraphs later, I wrote about the nature of principal mentoring programs and the type of design that might be best.

Mentoring is an experiential relationship. There are standards in requirements and techniques but also a focus on interpersonal relationships. This suggests a qualitative design to capture the characteristics of the mentoring relationship. But, to respond for the leadership field's more theoretical investigation into the effects of mentoring programs suggests a need for something more than just a qualitative design. Experimental models of evaluation do not seem to fit because the size of the program and the nature of the program do not fit a random control trial or quasi-experimental design. And, the resources of this evaluation are limited which prohibits an expensive and time intensive design. A purely qualitative constructivist design may give more of the same information that the literature seems to discount--embraces and documents context, but the generalizability of the results may be hard to justify.

A few more lines later, I came to this conclusion, "Adding constructivist elements to PTE would add contextual elements to the PTE, and would elevate context's importance to evaluation findings. Adding PTE to constructivism adds more structure and brings in outside viewpoints (e.g., social science and educational leadership literature) in interpreting the program and its effects on principals and schools."

In the end, I ended up designing an evaluation that used interviews, surveys, and document reviews to collect data on the program's context and to incorporate the perspectives of multiple groups of stakeholders. I also brought in research and literature from the educational leadership field to develop an a priori theory of how new principal mentoring program programs work to develop leadership skills and improve schools. The findings were analyzed and organized into a final program theory using the a priori program theory as the initial template with modifications based on the data collected through this evaluation.

Overall, all of the members of the metaevaluation panel rated my evaluation highly in terms of credibly constructing knowledge about the program. However, the evaluators on the panel said that I provided more data related to the program processes than the outcomes of the program. One evaluator said that the evaluation was "heavy on process and light on outcomes." I did a good job of identifying and describing the essential elements of the program (the, who, what, when, and where) but was not as successful in identifying under what conditions and to what effect.

The panel gave examples of data that they believed would have made the quality of the evaluation richer. First, the panelist wanted to see more information about the participants such as, the protégés' motivations for becoming a principal and their views on the role of the principal. Then the panelist said I should have probed into how the protégés' motivations and perceptions relate to their mentoring program experiences.

The panel also expressed a desire to learn about the outliers and non-responders to the survey to learn whether their experiences with the program were different from the responders, as well as the nature of the different responses to the program that the outliers may have had. The participants from the survey who provided responses outside the average responses may have

provided different perspectives that could have been probed for within the case studies and may have prevented the easy convergence that emerged during the interviews.

Program administrators on the panel said that they would have liked to have read more in-depth information about what happens in the mentoring sessions in terms of what they talk about, the qualities and characteristics of their discussion, and other mentoring strategies such as modeling, observations, and feedback, and the provision of resources. They claimed that it was difficult for them to make a link between the mentoring program and its outcomes without knowing what happens during the mentoring sessions.

Some of the comments I received from the panelists centered on the ambiguity of the testing and development of the final program theory represented in my evaluation. I had described to them the a priori program theory, but the panelists said that it was hard to distinguish between the development of the a priori program theory, the testing of that theory through my engagement with the program participants, and the final program theory presented in the report. The evaluators on the panel said they were unsure when and how I developed the theory, which one evaluator said is a common problem in PTE. One panelist said that I should have thought more about the origins of the theory in terms of who contributed to the theory and how and why it was shared (or not) in the district. Then I should have justified the design of the theory and examined how the theory worked under pressure with testing during the evaluation. A few of the evaluators said that the formulation of the a priori theory and the testing of it became too entangled and confusing. They recommended that I develop an a priori theory either with a literature review or develop a stakeholder-based a priori theory by engaging in the field, and then test the theory to explore a clash with how stakeholders experience the program and either refine or falsify the theory.

After reviewing the critiques I received from the metaevaluation panel with my review of the comparison evaluations, the contextual factors seemed to be a place where a lot of the evaluations fell short which impacted the evaluations' credibility. Overall however, the evaluations that used the realist, I-PTE and constructivist frameworks substantiated claims with evidence primarily using quotes and thick description from a triangulation of data sources from various stakeholder groups (e.g., surveys, interviews, observations, and document review).

The PTE of the managed health care system (Bickman, 1996) was particularly strong in a quantitative sense of developing an a priori program theory and using it to identify discrete outcomes and variables to guide statistical measurement that substantiated the claims that the program achieved its intended outcomes. However, no data were collected to make or substantiate claims about the quality of the effects the program achieved (e.g., increased follow-up doctor's appointments for treatment that led to higher quality of life outcomes for the clients). The evaluators also discerned those outcomes they would be able to measure and those outcomes that would remain untested in the program theory.

The interpretivist program theory evaluation (I-PTE) (Bowen, 1999) of a teen pregnancy program was another program evaluation that was particularly strong on creating strong linkages between the description of the program and the inferences made about the program. This evaluator used a mix of methods and bracketing methods to cross-check the findings of one data source against another. This evaluator also collected extensive contextual data that placed the program within its setting and helped the reader understand the program's essential elements and challenges. Similarly, the evaluator of the I-PTE of the Family Development Center (Crane, 2000) used a purposeful sample of those program participants who could provide information rich descriptions of their experiences with the program using interviews, observations, and focus

groups as primary data collection sources. Both of the I-PTE evaluators provided a graphical representation of their data in a program theory/logic model that was supported by rich description of the program theory using their data to substantiate the theories.

The realist evaluation of management practices in a Ghanaian hospital (Marchal, et al., 2010), and constructivist evaluation of an elementary school enrichment program (Ford, 1995) fell short in certain aspects of providing contextual data that helped the reader understand the development of some final findings. For example, as a result of the evaluation of the enrichment program (Ford, 1995), the district decided to eliminate the self-contained gifted classroom and infuse enrichment practices across the classrooms in the schools. This decision caused opposition by some community members and the teachers' union, but prior to that decision, no mention of this possible oppositional viewpoint was presented in the evaluation report. The evaluation report of the Ghanaian hospital (Marchal et. al., 2010) provided quotes and other evidence to substantiate their claims, but little contextual analysis of the demographics of the employees, size of the hospital's workforce, and so forth. Therefore, it was hard to discern when claims were made as to whether or not these findings were representative of the entire workforce or just the few who participated in the evaluation.

Fairness: Representing the multiple theories of stakeholders. When designing the evaluation, I made the choice to pick principals (current or former protégés) who had at least one year of experience in the mentoring program. I also aimed within my small sample of five principals to choose principals from elementary, middle, and secondary schools, a balance of male and female, and one African American principal. Program administrators from the district's central office, mentors who worked with the sample principals, and a broad survey of teachers

from the sample principals' schools were also tapped to provide input into the mentoring program's processes and outcomes.

While metaevaluation panelists said that I seemed to fairly represent the stakeholders most closely tied to the mentoring program and they could perceive no instances where I valued one group's data over another, several of the panelists said that they did not read much about possible different perspectives of the various stakeholder groups. Several of the evaluators on the panel wondered if the convergence was an accurate depiction of the stakeholders' perspectives or if my evaluation process was not sensitive enough to capture the differences. Another evaluator questioned whether the participants allowed me to lead them through the questions and answers in the interviews. One evaluator said that it would be possible that mentors and protégés could have had different perspectives about how the mentoring program works because of different motivations and agendas, but both could still agree that the program works. Because my evaluation did not probe into different motivations, agendas, and perspectives of a good principal, I was not able to pick up on these differences.

Looking at my reflection journal, about the halfway point in my interviews with program stakeholders, I found an entry in which I reflected on Schwandt's (1993) writing on the role of theory in inquiry in orienting us and engaging us in the field. I wrote:

The theoretical language that I used when designing the evaluation and interview questions relied heavily on the theoretical language and ideas used in the most recent educational leadership literature. I think that this helped me engage with the principals, mentors, and administrators because I shared their language. They readily seemed to understand my questions, although they had a harder time answering the outcomes-related questions more so than the process-oriented questions. I think that using this

accepted theoretical viewpoints also bought my credibility with the participants. From the way they nod when I ask the questions, I get a feeling that they feel that I know what I am talking about. Thus, I think this helped them open up. They didn't have to educate me on the finer points of leadership development, mentoring programs, and issues encountered by leaders because my questions led them to understand that I could identify with their perspectives.

With this entry, I did not come right out and say that I believed I was influencing or leading the stakeholders to say what I wanted them to say regarding the workings or the outcomes of this program. But, I wonder if that somehow through my language I did influence their responses and if the stakeholders picked up on something that I said and restated rather than expanded on my own preconceived thoughts about leadership and mentoring programs. In several entries after my interviews with stakeholders, I repeatedly complained about stakeholders' difficulty with describing to me how the mentoring program was impacting their leadership development, much less how the mentoring program was helping them to positively affect schools and student learning. Several times I mentioned feeling like I had to "pull" this information out of them.

Both evaluators and administrators on the metaevaluation panel agreed that the evaluation was fair in that different groups were invited to participate and that there did not seem to be anyone's perspectives that were held more prominently than others. However, there were a few groups that were possibly missing such as teachers and other district administrators who were not as intimately involved in the program. The panelists said it would have been interesting to have included the perspectives of members of the school board and community organizations who fund and are involved in education reforms in the district. One difference that these groups

might have described is how they define a good principal and a good school. These differences may cause them to think differently about how the mentoring program should work and if it is effective.

The findings of the comparative analysis found that overall all the evaluations covered the fairness criterion to a similar extent. All of the evaluations included multiple stakeholder groups using different methods as deemed appropriate for the evaluations' purposes. While my evaluation was critiqued for omitting the perspectives of teachers, school board and community members, and other district administrators, all of the evaluators of the comparison evaluations also made certain decisions to exclude key groups that may have offered different perspectives on the evaluands. For instance, Bickman's PTE (1996), included different stakeholder groups through chart reviews and billing records. Unfortunately, the data were not disaggregated by groups nor were in-depth interviews, focus groups or observations conducted to explore whether there were differences in experiences or outcomes for some groups (e.g., chronically ill patients, young children, adolescents). Similarly, the I-PTE of the Family Development Center (Crane, 2000) chose to sample from stakeholders who represented those participants with the most experience with the program rather than all stakeholders (i.e., those with a lot of program experience and those with only a little). It was a purposeful sampling strategy, but it may have added more to the evaluation to have the perspectives of those who had average or few experiences with the program. There may be a reason that some stakeholders did not engage with the program as much as others that may indicate a needed improvement or targeted strategy to make the program more inclusive. The constructivist evaluation of the Woodbury Elementary gifted program (Ford, 1995) did not have a good representation of parents from the school district to add a fuller picture of how parents (both those with students in the gifted program and

those whose children were not) felt about the program in the school. Interestingly enough, there was much discussion about the isolation of students in the gifted program and how often the gifted students were described as nerdy; but students were not included in the evaluation. The realistic evaluation of the hospital management services (Marchal et. al., 2010) appeared to include all of the appropriate stakeholder groups, but the sample sizes were small—sometimes only one person per stakeholder group. There was no explanation as to why the sample size was so small in relation to the total number of possible stakeholders in that group who could have participated in the evaluation. Finally, the I-PTE of the teen pregnancy program (Bowen, 1999) also adequately included the appropriate stakeholders, but two groups that were excluded from the evaluation would have added to the understanding of the program: parents of the pregnant and parenting teens, and fellow community service agency staff. Both groups were mentioned as either benefitting or challenging the program’s effectiveness, parents by not being supportive to their teen parents; and community service agency staff members were described as partners by teen pregnancy program staff but enemies by the teen parents. By including both of these groups in the evaluation, it would have been clearer how the parents and community agency staff members were affecting the lives of the pregnant or parenting teens, and program supports may have been modified to target the effects of these two groups.

Context program interdependence: Describing the relationship between the program and its context. Moderators in the PTE literature describe the contextual features that interact with the program and can affect whether or not the program reaches its intended goals. Dahler-Larsen (2001) called for a broader inclusion of constructivist principles into our program theory evaluations by exploring the bidirectional, interdependent relationship between a program’s context and the program itself. Often this relationship is represented as a

unidirectional relationship in which the contextual setting provides the impetus for a program and either benefits and/or challenges the program in achieving its goals. There was an overall consensus among all panelists that context is crucial to understanding how a program works and to be able to develop recommendations. In this evaluation, I described two moderators that I felt directly influenced the ability of the mentoring process to function effectively: the mentor/protégé matching process and characteristics of the mentoring relationship. However, panelists said that I should have capitalized the principles of constructivism by broadening this search for moderators that were not as obviously connected to the program. For instance, I included some information about the district philosophy and some of the reform strategies it was using to improve its schools and programs. I also recognized that the mentoring program was part of the professional development activities that the district offers to its administrators and teachers. However, both the evaluators and the program administrators felt that the description of the context was too narrow and lacked depth. I needed to capture the perspectives of a broader group of district stakeholders about the district philosophy and where the program fit in the larger district professional development and school reform initiatives. Exploring these contextual factors may have helped me capture contextual features of the district that helped and hindered the program.

The evaluators on the panel referred to Pawson and Tilley's (1997) realist evaluation in which the evaluators would look at the motivations and world views of participants and consider how they match up (or not) with the program philosophy. For example, one contextual feature that may impact the effectiveness of the program could be the motivating factors of the new principals. Why did they become principals in the first place and do they believe that they have the capacity in their role to improve the school? Why is it helpful to talk to a mentor? What are

the issues they are dealing with in the job? How are they handling the work-life balance? What is their status in the school, the district, and the community? These are questions that could have been explored and used to develop a hypothesis that could have tested in the evaluation. Another district factor that could have had a significant effect on the program is the district's impetus for implementing the program. What problem were they trying to solve, and did the various stakeholders in the district share the view that this was the right problem and the right solution to that problem? One panelist said that it is a common problem with PTE that evaluators begin the evaluation assuming that the program is a solution to a problem and they slip into a deficit model. But, this panelist wondered if the district really had a problem with poorly prepared principals who were struggling in the first years of their principalship.

Program administrators on the panel focused on the need for exploration of school-based contextual features. They wanted to read about differences in the district from school to school and how these differences may have affected the mentoring program. One example is the effect that teacher union contracts can have on mentoring. Notably the culture of a school that is in a combative situation with its union may have a different influence on the mentoring program than the culture of a school in which the union relationships are stable. The mentoring process can look quite different for a principal who has the support of the union to implement the changes versus the mentoring process for a new principal who has a union that is resistant.

A continuing strand of journal entries straddles both this category of findings and the next—the context program interactions and identifying program effects, respectively. In these entries, I wrote about the relationship between the mentoring program and other professional development programs and education reform initiatives in the district. Early on in the interviews, it became evident that the district administrators who put this program into place did so as part of

a set of leadership reform strategies without much forethought about the goals and desired outcomes specific to the mentoring program. Within a couple years, they began to appreciate the mentoring program as a support mechanism for other professional development and education reforms and strategies they had adopted for the district. Mentors and protégés stressed the benefit that they could attend professional development events together and then discuss them afterwards. For instance, in one entry in the reflection journal, I wrote how two program administrators told me that the district's professional development is focused on three specific leadership foci. Because the mentoring program is under their purview as the human resource administrators, the mentoring program is also focused on these three foci. Thus, the mentoring program helped new principals pull all of what they learn together. It was at that point that I began to see the mentoring program as a support mechanism for other professional development that may be much more explicitly focused on improving instruction and student learning. While I began to make note of the supportive mechanism of the mentoring program in the interviews, it is an area that I should have explored more with a broader set of district stakeholders to more fully understand this interdependent relationship between the different components of the entire system of the districts' efforts to improve leaders and schools. I write about how this relates to identifying program outcomes in the section that discusses the findings for the utility criteria.

While my evaluation received critiques for not addressing the context program interdependence as broadly as I should have, I believed I did try to make a concerted effort to understand the relationship between the mentoring program and how it was situated within the districts' professional development activities for principals in the school district. From the interviews with program stakeholders it became evident that the mentoring program was a support mechanism for supplementing and individualizing the topics addressed in the whole

group professional development sessions. Mentors and protégés would take those topics back to the mentoring sessions, expand on the topics, and tailor the topic to fit the needs of the protégé and the school. This, in turn, the mentors and protégés believed would help the protégés be more successful in their schools.

The evaluations in the comparison sample were less strategic in exploring the interdependent, bidirectional relationship between context and programs. As described above, all of the evaluations (excluding Bickman's PTE), provided extensive description of the contextual features of the programs under study. These descriptions included demographic information of the targeted populations served and program personnel, organizational characteristics (e.g., organizational charts, organizational policies and practices), descriptions of the targeted problem the program addressed, and challenges to the program's implementation and effectiveness. However, these evaluations did not address how the program itself acted on the context itself. For example, the I-PTE of the teen pregnancy program (Bowen, 1999) could have brought in the perspectives of the parents of the teens and community service agencies to better explore how the program affected their perspectives on the teen parents. One of the issues that most concerned staff members in the teen parenting program was that many of the program's desired outcomes—delayed second pregnancy, high school graduation, college attendance, etc.—were not being attained. This program evaluation did uncover an implicit goal of the program staff of developing a trusting relationship with the teen that they believed most of the teens did not have in other relationships in their lives, namely with their parents and other community service agency personnel. Staff believed their trusting relationships with teens would eventually lead to the teens' self-sufficiency and less welfare independence. Another avenue of exploration would be whether or not the effects of the trusting relationships translate into the teens behaving

differently with their parents and service personnel which may lead to better relationships with these people as well. Or, if parents and service personnel change their opinions of the teen parents knowing that they were taking the initiative to participate in the program, thus indicating they were taking more control of their lives, which would lead to better outcomes. There is also the chance that the teens become overly attached to and trusting of the teen pregnancy personnel that they become more reliant on their help rather than reaching out to parents and service personnel to work on building those relationships.

Of the five evaluations in the comparative sample, the constructivist evaluation of the enrichment program (Ford, 1995) treated the interdependency between the program and the context the best. Ford discussed the contextual setting within which the program resided in the elementary schools as a self-contained classroom for gifted students. As a self-contained classroom, the teachers and students were primarily isolated from the rest of the teachers and students in the schools. Parents, teachers, and students felt confused about the program and believed that it was elitist. The selection criteria seemed to be unduly weighted on test scores and not on student motivation or other characteristics not well measured on a standardized scholastic test. Others in favor of the program saw the benefits to gifted students who received special learning opportunities (e.g., field trips and projects) that motivated them to achieve and expand their achievement. Those stakeholder beliefs about the program either acted on the program as benefits (from those in support) and challenges (from those in opposition). Conversely, the evaluators described the impact of the program on the context, largely in terms of what was absent. The original intent of the program was that the enrichment teachers would provide professional development to those in the general classrooms to model and instruct on enrichment program learning strategies. The hope was that these strategies would improve the teaching and

learning in classrooms across the school rather than just the self-contained classroom, but this did not come into fruition. Instead, the teachers had a segregated system whereby teachers in the self-contained classrooms kept to themselves.

Principled discovery and competitive elaboration: Studying the program and identifying the program's effects. While the panel said that I constructed credible knowledge about the program, represented stakeholders fairly, and provided adequate descriptions of the relationship between the program and its context, my evaluation broke down in providing evidence of the program's effects. In PTE, the effects are included in the model as mediators and outcomes.

In terms of mediators, I identified a few proximal outcomes that I believed mediated the attainment of short-term outcomes. The two mediators I identified were an improvement in the new principals' confidence in their role and improved adjustment to the position. However, a few of the evaluators on the panel questioned if improved confidence and adjustment really does lead to improved leadership knowledge and skills. In a few journal entries midway through the interviews, I reflected on how to represent the relationship between the two mediators and the short-term outcomes because the way this relationship was presented in the program theory seemed too linear when the actual relationship seemed more circular and interdependent. For instance, with increased confidence and adjustment to the position as leader of the school, these new principals seemed more apt to branch out to perform leadership tasks they believed important to improving the school, but that may also be difficult to accomplish--whether politically difficult or technically difficult to implement. However, the more that they were able to perform these leadership tasks, the more confident and adjusted the new principals became, and hence more likely to perform needed leadership tasks. When they felt they had failed in their

tasks, they could rely on their mentor to help them re-strategize their practices and bolster their confidence. Hence, this was a circular relationship rather than the more linear relationship that was depicted in the final program theory. Presenting the relationship between the mediators and outcomes as a circular relationship rather than a linear relationship may have helped the evaluators on my metaevaluation panel think about the effects of the mediators on the short-term outcome of improved leadership practice.

After the mediators in the program theory, I identified short-term outcomes that relate to protégés' improved leadership knowledge and skills. There was not any objection to these outcomes, but there were questions about the data used as evidence to support them, more specifically the methods used to collect that data. First, the evaluators on the panel were concerned that most of the data were self-reported by protégés in the surveys and the interviews. They said there was the possibility that mentors and protégés liked the program and that may have biased their responses to spin a positive light on the program. Both evaluators and administrators said that triangulation of these data with data from more objective observers such as teachers or observations of protégés leadership knowledge and skills by the program evaluator or administrator/supervisor would have added validity to the inferences. Teachers and others in the building may have also been able to provide data on what effect protégés had, or were having, on the school climate and curriculum and teaching in the building.

To improve the validity of the evaluation influences, panelists suggested studying this program using a comparative case study that asked what would have happened in principals' development without this program. One panelist suggested a quasi-experimental study of mentoring program participants compared to a similar group of principals who did not participate in the mentoring program. I contemplated doing this when designing the evaluation, but due to

time and resource constraints, I was limited in my ability to find a comparison group outside of the district. Per my constructivist viewpoint, I was also more interested in digging deeply into this district's program and using district data to study the program.

Metaevaluation panelists also said that my evaluation did not rule out alternative explanations to account for the positive findings that I attributed to the mentoring program. They asked how the mentoring program compared in its effects to other leadership development activities and other education reform efforts in improving schools. Program administrators said I should have probed into what protégés valued more—the mentoring program or other professional development activities such as graduate program or coursework, in-services or principal professional development academies, or other mentoring opportunities outside of the program or district. It is possible that the protégés who participated in this evaluation also may have participated in other professional development activities that they valued more and credited more for developing their skills. Finally, one evaluator asked if I had found that there were no differences in student learning in protégés' schools, could an alternative explanation be that the mentoring program is a distraction for the principals that kept them from doing the things that really lead to student achievement.

Throughout this evaluation, I continually weighed the evidence for the effects of the mentoring program against the possible effects of other professional development and reform strategies. In one entry, I wrote

In my evaluation, I built an a priori theory of how and why mentoring programs work that helped guide the design, implementation, and analysis of the evaluation findings.

This provided a structure for investigating the program--its impetus, goals, processes, and

outcomes. However, it might also have led me to turn a blind eye to some possible alternative explanations [for the results obtained].

The more I talked with stakeholders, the less able I was to tease out the effects of the mentoring program from the effects of these other leadership supports in this district. Furthermore, I did not think that the effects should have been teased out because I came to believe that serving as a support mechanism to these other efforts may be the mentoring program's strongest effect. Going back to the interdependency of the professional development and reform initiatives and the mentoring program, there were several entries where I struggled with how to represent this interdependency and its effects in the program theory, whether it belonged in the program's chain of effects or if it was considered part of the contextual relationship. I ended up putting this interdependency in the description and graphical representation of the context. However, I still question if that was the correct decision, and if maybe there should have been a complementary program theory that further illustrated the supportive role the mentoring program played in the district.

The panelists considered whether the protégés who reported success with this program would have succeeded regardless of this program. Others wondered if those who were disgruntled with the program chose not to participate in the study, so that my findings were skewed because only those who had positive experiences participated. An evaluator on the panel suggested that I "follow the money." This mentoring program was part of a multi-million dollar initiative, and he wondered if the protégés received other resources that they found just as valuable as the mentoring which may have influenced their interest in the program regardless of whether it caused improvements to their leadership knowledge and skills. For instance, these new principals (i.e., protégés in the mentoring program) may also have liked the extra attention

they received from the program such as extra resources, networking and opportunities to participate in a learning community with other principals, and dialoguing with another person whether or not it helped build their knowledge and skills.

Finally, one evaluator was concerned about the notion of extending the program theory all the way to student learning. He questioned who wanted such a long causal chain and why it was not enough that the program would improve the principals' skills and lead to an improved school culture. A few of the administrators also said that it is hard enough to demonstrate that the mentoring program improves principals' knowledge and skills much less student learning. The evaluator said that identifying the outcomes can be interesting and politically controversial, but as the evaluator we must have this conversation with stakeholders and then be ready to justify and test the outcomes included in the theory. My journal entries revealed my recognition that the ability to identify effects on student learning was problematic from the beginning of the evaluation. With each interview, I noted the difficulty that the program participants had in being able to identify effects on students, and seemed confused by this line of questioning. After one interview, I wrote:

The answers to the questions came after considerable probing...I felt [name of principal] had trouble understanding some of my questions or how to answer them, especially when I asked how the mentoring program helped her change her leadership practices and the outcomes at her schools...The more distant the effects from the program, the harder it was to explain how the mentoring program was helping her affect student outcomes.

I suspected that if I had asked an open-ended question about the outcomes of the mentoring program, student learning outcomes would not have been mentioned, except in passing that ultimately the goal of any reform or school improvement strategy is to improve

student learning. After data collection had ended and I was analyzing the data, I came to this conclusion about program outcomes:

There is a constructivist notion that the objectives of a program are negotiated between program actors. The [name of district] program theory supports this notion. I don't think that there are prescribed outcomes. Each mentor relationship is based on the individualistic needs of the new principal, but due to the varying a priori experiences and knowledge of the new principals, the level of support offered varies from relationship to relationship. The ultimate goal for all new principals and the school district the ultimate outcome is increased student achievement. However, the goals each new principal wants to focus on to get to that ultimate outcome differs across protégés.

Through the comparative analysis, it appeared that the different evaluators identified and measured outcomes differently from my evaluation. Bickman (1996) chose the outcomes based on the program theory, identified measurable variables, and used appropriate statistical procedures to collect the data on those outcomes. The realist evaluation (Marchal et.al., 2010) used a similar procedure albeit on a smaller scale than Bickman's and using a mix of methods. The I-PTE's and constructivist evaluations, on the other hand, had little to no focus on outcomes. The I-PTE's were primarily focused on uncovering the program theory that guided successful implementation of the program, and the constructivist evaluation told the story of the district's and community's perceptions of the enrichment program that led up to a final decision on the fate of that program.

The competitive elaboration criterion, as I interpreted it and used it in this study, dealt with eliminating alternative explanations to the attributions made about program implementation and effects. While constructivist principles expect and advocate for multiple theories of program

functioning and effects, representing multiple theories in a PTE approach can often be difficult. Often, programs represented through a PTE are reported as multiple perspectives eventually converged into one program theory. The realist evaluators and constructivist/interpretivist evaluators allowed for a more emergent and multiplistic account of the programs. The realist evaluators of the Ghanaian hospital (Marchal et. al., 2010) developed multiple context-mechanism-outcome models to account for the diversity of theories about how the management principles were implemented. The I-PTE's of the teen program (Bowen, 1999) and family development center (Crane, 2000) focused on the implementation theories and whether the programs were successfully implemented. The critique for these evaluations was not that they did not rule out alternative explanations but that they did not explore alternative explanations as thoroughly as they could have as mentioned before in the sections on context program interdependency and fairness. For example, by excluding parents and community service personnel from the evaluation, Bowen (1999) missed an opportunity to explain how the program impacted these stakeholders' relationships with the teens that could give a fuller understanding of the program's impacts. Similarly, Ford (1995) would have given the reader a stronger explanation of why the enrichment program in self-contained classrooms was abandoned in favor of enrichment time in every elementary classroom and the potential effects on students and teachers in the schools. Therefore, while from the PTE perspective, the critique may be that evaluators did not rule out alternative explanations, from the constructivist perspective, a critique would be that the alternative explanations were not more fully explored and described in the evaluation reports.

Utility of a Constructivist Program Theory Evaluation

This section highlights the findings from the analysis of the reflection journal and metaevaluation panel responses to questions about the utility criteria. The criteria included in this analysis are: resonance, transferability, information scope, and values identification. The utility criterion of report clarity was not analyzed for this study because this criterion was used in the metaevaluation panel to be sure that the panelists' responses and perspectives about my evaluation were not due to a badly written report. It was not a criterion used for testing the utility of my evaluation for the sake of the metaevaluation.

Resonance: Understanding the program. Overall, the panelists said that after reading the evaluation report they had a good sense of what was happening in the mentoring program. Program administrators said that while the mentoring program's policies and structures were somewhat different from those of their own programs, they still saw many similarities and read the report through their own lenses of how these programs work.

However, the panelists said there were some important pieces of information that might have helped them understand how and why this program works (or not). First, the evaluators reiterated their concern about the lack of causal explanations that would have helped understand if this program works. A comparative analysis that identified essential elements of the program would have helped. Second, one of the evaluators said the lack of a link from the program to changes in student achievement was problematic, although he had some issues with including student achievement into the program theory in the first place. While he understood that student achievement is often included because of accountability pressures from government agencies, it should be understood that as evaluators we must be ready to test and defend it. As evaluators we need to ask whether improving student achievement should be the purpose of the program

anyway because principals have such an indirect effect on students and generally the purpose of mentoring programs is to improve the knowledge and skills of principals. Case descriptions that described the patterns of effects from effects on protégés to effects on school climate, and ultimately on students would have helped readers gain a fuller description of how and why this program works.

While program administrators said that despite differences in policies and structures from their own programs they could understand the nuances of this program, there were a few elements missing from the description that would have helped them better understand the program. First, the report did not describe a formal screening process for choosing mentors. They said this is important because a good principal does not always make a good mentor and just because you know and like a person, does not assure he/she will be a good mentor either. Second, administrators did not understand the charge to mentors. What was expected of them? Were they supposed to set expectations for what protégés would learn or achieve during the program and then collect evidence of it? Finally, administrators said that the report was missing a discussion about whether or not the mentoring program was addressing the discrete skills and knowledge principals need to be effective. So, a next step for the district may be to formulate a tighter description of what it is to be an effective principal in the district, and then to make sure that the school board, mentor, and program are on board with that description.

The analysis of my reflection journal found that one of my primary concerns was developing a model that adequately addressed the values that stakeholders held in relation to this program. One interesting note was that none of the metaevaluation panel noted the lack of stakeholder values in the program theory/logic model. It was evident as the interviews with administrators, mentors, and protégés that they highly valued the extra support and nearly

instantaneous feedback the new principals received through their mentors off whom they could bounce ideas. The stakeholders valued the mentors' accessibility to individualize the professional development that new principals received as part of the district-wide principals' group meetings. The protégés also valued being part of a network of new principals that gave them another feedback mechanism that helped ease the isolation and adjustment to the principalship. All of the protégés, past and current, recommended that the mentoring program continue. Hence, this program was highly valued. My struggle was in how to represent these values in the model. The values were, in a sense, the outcomes. The protégés valued these support mechanisms as a result of participating in the program; the protégés may have valued support, professional development, and networking outside of the program. Largely though, the values of these concepts were localized and solidified as a result of participating in the mentoring program. Therefore, I did not place these values in the contextual description of the program setting or in describing the stakeholders.

I also struggled with representing and describing the complexity and interdependency of the different mentoring program components and effects. In the model, I placed improved confidence in the role and a smoother adjustment to the principalship as two mediators that were immediate outcomes of the program. They were the two effects that I heard repeatedly from program administrators, mentors, and protégés. The evaluators on the panel questioned whether these two mediators did lead to improved leadership practices. In listening to the stakeholders describe the effects of improved confidence and a smoother adjustment, it appeared that these two mediators positively affected the protégés' abilities to implement appropriate and successful leadership practices which then improved their confidence and eased their adjustment to an even

greater extent. Therefore, the relationship between the mediators and the short-term outcome of improved leadership practice was circular and compounding rather than linear.

At several different points of the evaluation, I wrote in my journal about the complexity of causality, particularly within school settings. Here is one entry:

[H]ow we as educators, parents, researchers, and community members think about how we can impact student achievement is a highly complex and cumulative series of social and individual actions and contexts. Students interact within a system that affects whether or not they learn. This system includes family characteristics (e.g., family income, health, education level), community characteristics (e.g., poverty level, community resources, rate of violence, employment opportunities), and also school characteristics (e.g., teacher and leader effectiveness, curriculum, school culture, resources). Research has not identified that one thing that leads to improved student outcomes as there is not one thing. Therefore, can we really isolate student learning as an effect of a mentoring program on a principal's performance in the school? If not, how do we represent effects on student learning as principal performance is just one variable in a constellation of variables?

So, I agreed with the panelists, the causal linkages were lacking in my evaluation, but like Bickman (1996) I tried to represent the untested outcomes with arrows with broken lines. However, I was more bothered by the simplicity with which I felt the PTE dealt with outcomes because for each of the short- medium- and long-term outcomes, there could be another constellation of program theories and contextual analyses to more fully explore. The program administrators on the panel seemed to be aware of the problems with tying program effects to student achievement because one panelist said that it was hard enough proving effects on leadership practice much less student learning.

Looking across the evaluations in the comparison sample, the evaluations conducted within the I-PTE and constructivist models had the strongest descriptions of the program. From these evaluation reports, I felt that I understood the complexity of the program processes and the program context. The program theories of the programs developed for the I-PTE's (Bowen, 1999; Crane 2000) provided the best understanding as the graphical representations grounded the rich description and narrative and showed the links between program processes and the program's effects. The constructivist evaluation (Ford, 1995) provided the story of the evaluation process and resulting actions to eliminate the self-contained gifted classroom. Interestingly enough, the evaluation did not include a rich description of the enrichment programs (e.g., instructional strategies, learning environment) so I was not sure of how these classrooms differed from the general classrooms. In reading the report, there also was not much reference to the deliberation around eliminating the self-contained enrichment classrooms. Therefore, it was a surprise when that was announced at the end of the report and difficult to see what led to that ultimate decision. It did not seem that other alternatives were given much reflection.

The realist evaluation (Marchal et. al., 2010) and PTE (Bickman, 1996) did not provide an adequate description of the program or policy that they evaluated or the descriptions of the community or population served. I walked away from these evaluation reports not knowing what the management practices or managed healthcare policies looked like beyond a cursory description that gave me the gist of their purposes and overarching practices. However, Bickman (1996) provided a clear description of causality in the narrative and program theory model. He used coded arrows to indicate the strength of the linkage between program policy and the effects in which bolded arrows showed a strong statistically significant effect, and arrows with broken lines were causal linkages that were untested. Marchal et. al. (2010) did not present the causal

linkages as well. The evaluators described in some sense the different Context-Mechanism-Outcome models developed out of the evaluation, but they were not explained in much detail, nor were they presented in a graphical or table format that would enable investigation by the reader.

Transferability and information scope: Useful knowledge for informing the design and administration of programs. Program administrators answered questions about whether or not the evaluation findings related to their own experiences with mentoring programs and if the information would inform their design and administration of their programs. As I wrote above, the Clearlake mentoring program had different policies and structures from the programs these administrators manage such that the Clearlake program was an in-house mentoring program where the administration, mentors, and protégés were all from within the district. The administrators on this panel managed programs that outsourced mentors to several districts and matched mentors with protégés who were not employed in the same district. With those differences in mind, administrators said that the findings inspired them to re-think some issues related to mentoring.

In the Clearlake program, all of the mentors were either practicing or recently retired principals from within the district. However, administrators on the panel oversaw programs in which mentors were recruited outside of the district. On the one hand, they could see a benefit for the district in that using in-house mentors could allow a district to embed certain beliefs about leadership and their expectations for what it is to be an effective leader in that district. However, the administrators were also cautious and highlighted problems in-house mentoring could pose. First, mentoring could become a buddy relationship that focuses more on getting the work of the district done than on developing a leader for the district. Mentors and protégés may use meetings

of convenience (e.g., during professional development in-services) rather than setting aside focused time for mentoring. The district may also face a problem of confidentiality if the mentors and protégés became friends, or if one day the mentor would become a past protégés' supervisor and would know certain things about these principals (e.g., leadership weaknesses) that could hurt them in the future.

There were certain mentoring strategies that administrators rethought. The first is whether or not mentoring is the same regardless of level of schooling (elementary, middle, high school). In other words, should mentors and protégés have experience in the same school levels? They said that the discussion was not clear in the evaluation report. Even so, one administrator drew on her own experience and stated that mentors and protégés should come from the same school level because the gap in issues between an elementary and high school are too large. For example, high school principals deal with certain liability issues that elementary school principals do not (e.g., sports participation and events). The second strategy that panel administrators rethought was the use of modeling in which the mentor either invites the protégé into his/her building to model a leadership strategy, or the mentor goes into the protégé's building and demonstrates a leadership practice with faculty, staff, or students. Administrators on the panel said that in their experience, modeling works better when the mentor and protégé visit a third school where another principal is effectively practicing a leadership strategy that the mentor or protégé would like to see in action. Otherwise, the modeling and the mentoring relationship can become all about showcasing the mentor rather than developing the new school principal.

While recognizing that the Clearlake mentoring program was an in-house program compared to their own programs that outsource mentors to a variety of districts, the

administrators wondered how well the findings generalized to their programs. Despite this structural difference, administrators said that there were several findings that caused them to think about their programs and practices in a different or deeper way. For instance, the administrators asked themselves how a district supports a principal mentoring program and institutionalizes it. How does a mentoring program fit with the district's overall professional development plan and programming? What role would the school board have to play to support and institutionalize the mentoring program?

Within the mentoring program, administrators have to develop mechanisms and coordinate the processes of the program such as content, mentor selection, and the mentor-protégé matching process. Coordinators of principal mentoring programs have to think about how the mentors will help the new principals develop into skillful leaders. To do that, the panel administrators reflected on the content of their programs and how to present technical issues primary to the principal position and critical events in the lifecycle of schools such as staff cuts, hiring, professional development, and dealing with tricky staff issues. There were content pieces discussed in the evaluation report, but administrators said that they would have wanted to see some of the content fleshed out more. One noted examples was the discussion on the district's focus on a set of leadership strategies nicknamed by the district as the "Three Buckets" (Leithwood et al., 2005) that guided much of the district's professional development programming. The panel administrators did find the discussion on the critical events that guided the timing of content (e.g., staff cuts and hiring in spring) to be helpful. To them it seemed intuitive and practical to focus the mentor and protégé on leadership issues in a "just-in-time" fashion at the point in time when the new principal was dealing with these issues.

The second program process that affiliated closely with the panel administrators' work was the matching process. From the evaluation report, the administrators noticed that the district did not appear to have a formal mentor-protégé matching process. It seemed to be based on the administrators' knowledge of personalities, characteristics of the students both mentors and protégés each served, and new principal leadership strengths and weaknesses. In response to what appeared to be more informal criteria, administrators on the panel said they believed that it is important in mentoring programs to have a more formal matching process based on specific criteria. One administrator noted that in a regional meeting of principal mentoring program providers, a meeting participant said that the matching process was not important. But this administrator said that was because so far the process that the participant was using was effective so they were not experiencing any mismatch issues. But, if the matching process does not go well and there are mismatch issues, then the importance of the matching will become evident. Another administrator on the panel said that the discussion about the matching process in the evaluation report caused her to think about what are the variables that she looks at when doing the matches, like school level, gender, and ethnicity, and others.

The third program process on which administrators reflected was the mentor selection process. Based on the evaluation findings and the recommendation that program administrators should consider the length of time a principal has been retired when choosing them as a mentor, the administrators on the panel said that based on their experience it was not the length of time that was important but what they were doing during that time out. So, according to the administrators, a retired principal who was continuing to remain engaged with the leadership and education field (e.g., continuing professional development, attending conferences) would be a better mentor than one who disengaged from the profession and education field. One limitation

that program administrators often have to navigate is that superintendents are usually reluctant to release principals to be mentors because of the amount of time spent out of the school. In the case of the Clearlake District, the superintendent implemented the program and saw the value, so release time was not as much of a barrier for the district as it could be for other districts with less district administration commitment. But in cases in which the program used external mentors (i.e., external to the district), the administrators had to rely on retired principals who do not have that limitation. Therefore, they recommended that mentor selection should be on a case-by-case basis with some flexibility in the criteria that deal with the amount of time in retirement.

Values identification: Criteria for judging program quality. Panelists were asked if the criteria used to make value judgments about the mentoring program were clearly identified and appropriate. The criteria used for the evaluation of the Clearlake mentoring program are included in Appendix J. The data on these issues were mixed. Overall, the panelists said that the criteria were evident but still offered some challenges. The evaluators on the panel had many more issues with the criteria than did the program administrators.

The first set of issues the evaluators had was in regard to the origins of the criteria. Panel evaluators said that they were superficial and haphazardly overlaid in the evaluation design. It appeared as if they were based on research and had been identified as precursors to success. Because the criteria seemed to come out of a publication written by program officers from the foundation that funded the mentoring program, the evaluators wondered if that introduced bias into the judgments of worth and success. Furthermore, the adoption of criteria from an external source is in opposition to a constructivist stance, and I needed to have provided a technical justification for including these criteria (e.g., criteria are correlates of success).

Instead, the evaluators suggested different ways to develop criteria for a C-PTE. First, criteria could be developed out of the model (e.g., an increase of 25% on a school climate survey). They also suggested that I develop criteria based on the particular focus of the evaluation whether it is looking at process or effects. Thus, criteria related to process would examine the quality of the implementation and the fidelity of implementation. If the evaluation is also examining effects, then the evaluation should include criteria related to expected outcomes (e.g., the 20% increase in school climate surveys as mentioned above) and demonstrate the relationship between criteria and principal and student outcomes. After developing the criteria, they suggested I get feedback from the stakeholders to check whether or not the criteria and levels of desired effects are acceptable. However, one evaluator said that gathering feedback from multiple stakeholders could get complicated if stakeholders hold different views about quality, worth, and effectiveness. For instance, criteria related to how well the program affects principal leadership skills could become an issue if stakeholders hold different definitions of a good principal and the skills the principal should develop and enact. PTE would have some difficulty in portraying these different definitions as the natural inclination is to bring convergence of perspectives into one program theory. Constructivist principles would expand upon the differing definitions and explore how each of these definitions impacted program implementation and the mentors' and protégés' experiences with the mentoring program. Depending on the definition, mentor and protégé may have different experiences with the mentoring program. If mentor and protégé had differing definitions of a good school or a good principal, one could imagine that these competing definitions could impact the mentoring/protégé relationship, as the trust and rapport that are important to the success of the program may never be built. In the mentoring sessions, the mentor or protégé may see certain

goals or strategies as being an important focus, but if the other does not agree that those goals or strategies are important from their differing point of view, the mentoring sessions may become tension filled and unproductive.

Overall, program administrators on the panel were generally positive about the criteria used in the report. However, they did make a few interesting comments about the criteria related to program effectiveness. First, one administrator said she was not clear on how the criteria addressed program effectiveness. She said that she based her judgments on what the program was supposed to do, on the program's goals and purpose. Another administrator was pleased with the criteria and the report recommendations and said the recommendations were "gold plated."

In looking back at my journal entries, I found that this was an important tension in this evaluation. What criteria to use when conducting a C-PTE? I think that it is actually one of the places where the constructivism and PTE can easily integrate because both have criteria that should grow organically out of the evaluand. PTE also allows the introduction of external sources of criteria from research and other literature associated with the field of study. However, there are also tensions. For example, PTE aims for parsimony and data reduction. Constructivism values expansion and rich description of complex programs. PTE values explanation and prediction; constructivism values understanding and highlighting patterns and connections. Therefore, identifying criteria for judging program quality was difficult. I knew that I would not have data that would show, for example, a 20% increase in XYZ practice. The goals for the program were nebulous as administrators, mentors, and protégés had difficulty identifying the goals and desired outcomes of the project. Therefore, I felt like it had become my responsibility to identify a set of criteria and chose a set of criteria that was external to the program and derived

from a scan of the mentoring program literature. As noted above, my undertaking of this identifying the criteria was problematic for the evaluators and one of the program administrators on the panel as they believed that the criteria should have developed out of the program's goals.

The comparative analysis showed that not one of the evaluations in the samples clearly identified the criteria by which programs were deemed either successfully implemented or effective. It was the assumption that for the PTE (Bickman, 1996) the criteria were whether or not the analysis reported statistically significant improvements in the healthcare treatment of children in military families. I assumed the goals of the programs under study in the realist evaluation of hospital practices (Marchal, 2010) and the two I-PTE's (Bowen, 1999; Crane, 2000) served as the criteria for successful program implementation. Finally, the constructivist evaluation (Ford, 1995) did not explicate criteria either. The final decision on the enrichment program seemed to be based on whether or not the problems with the enrichment program (e.g., non-diffusion of enrichment practices to the general classrooms and confusing selection criteria) could be solved. Because they could not be solved, the self-contained enrichment program was disbanded and enrichment was expanded to all classrooms. For all of these programs, I, the reader had to assume these were the criteria to judge the overall merit or worth of the programs, but it is feasible that my interpretations could be incorrect.

Summary of the Results

This chapter presented the findings of my analyses from three data sources: my reflection journal that contained my thoughts as I designed and implemented the mentoring program evaluation, the responses from a metaevaluation panel that consisted of experts in the evaluation field and mentoring program administrators, and a comparative analysis of five evaluations conducted with the PTE, I-PTE, realist, and constructivist models of evaluation. In general, I

found that my evaluation had both its strengths and weaknesses. My primary strengths were that the evaluation provided a good contextual understanding of the mentoring program and fairly represented the multiple stakeholders, and the program theory structure of the evaluation helped the reader understand the program's implementation. In addition, mentoring program administrators felt that they learned about different aspects of mentoring programs that made them think differently about the mentoring programs they administered and that there were some strategies that my evaluand used that they might find helpful to improve their programs.

The weaknesses in my program evaluation were that I was not broad enough in my contextual exploration of the program in that I remained too close to the mentoring program and did not broaden my search for contextual features further away from the program that may have affected its implementation. In addition, the constructivist principle of representing multiple program theories may have been largely violated as the multiple perspectives of stakeholders too easily converged into one program theory. For the most part, in comparison to the sample evaluations, these strengths and weakness played out similarly in them as well. In the next section, I present a thematic analysis of my findings across all data sources.

Thematic Analysis

Theme 1: Context. Context became an even bigger issue than I first envisioned. In the constellation of my findings, context would be located in the center of my findings with all other findings orbiting around this issue. As I presented my findings, I began with an entry from my reflection journal about how I wanted to honor the complexity of the mentoring program as it is situated with a highly complex education system. I blended constructivist and program theory evaluation so that I could bring in the depth of experience that constructivism values and the structure of program theory evaluation that would reign in and focus the multiplicity of

constructivism. From reflecting on the critiques given to me by my metaevaluation panel, my thoughts in my reflection journal and the comparisons with other evaluations, it appears that context poses a very difficult set of decisions for evaluations. First, there is the tension between depth and breadth. The evaluators on my panel were more focused on the characteristics of the individuals, while program administrators were more interested in school level characteristics. It could be that program administrators (many of whom were former principals or other school administrators themselves) were not as interested in the individual principals' motivations or agendas, because they were more interested in whether the principals had the knowledge and skills to serve the needs of the school and district. Therefore, school and district contextual features mattered more.

Many descriptions of the context of the Clearlake School District within which the principal mentoring program resided dealt with the local climate that surrounded the program in terms of local issues or politics, local demographics of the locale and targeted population, and the definition of the problem. Not one of the evaluations (in the comparison sample or my evaluation) dealt with the context program interdependency criterion well. All of the evaluations provided good descriptions of the context in the terms just described and how that context set the stage for the program under evaluation, but not one evaluation described the interdependency even though it seemed obvious that an interdependency probably existed (e.g., the community service agency personnel's possible negative attitudes toward the teen parents). It may be because by trying to uncover the interdependency, the possibility of ruling out alternative explanations for program results would become much harder.

The evaluators on the panel were interested in the motivations and agendas of both the protégés and mentors in the program as additional features of the program's context. They

believed that these motivations and agendas may have had an impact on the functioning and perceived effectiveness of the mentoring program. My evaluation was not the only one to have ignored the motivations of program stakeholders; the other evaluations in my comparison sample also excluded these features from their analyses. As I stated in the preceding paragraph, contextual features tended to be those external features that are listed almost as a checklist: locale description and demographics, stakeholder demographics, problem definition, and program process description. Unfortunately, the values, motivations, and belief systems of stakeholders are often not included in our descriptions of the context, nor are they considered in how they impact how stakeholders experience the program or the program's effectiveness.

By not dealing well with contextual issues, we hurt the quality and utility of our evaluations. A continuing theme throughout my reflective journal and in the metaevaluation was that context often compounded the difficulty in understanding how and why programs work or do not work. When critiquing my evaluation on the quality and utility criteria, the metaevaluation panel continually came back to breakdown in how I missed important contextual elements—mostly that I did not go deep enough in exploring the district, school, program and stakeholder characteristics that may have helped them understand the functioning of the program better and provide stronger, more credible linkages between the program and its outcomes. They also suggested that by going deeper, the outcomes and criteria for judging program quality and effectiveness may have been different from the outcomes identified in the mentoring program theory. It is my belief now that the preliminary outcomes that I identified as mediators—increased confidence and adjustment to the leadership role—would have been the primary outcomes associated with this program. I also would have ceded that because of the interdependency of the many variables that are brought to bear on the effectiveness of a principal

(e.g., past leadership experiences, additional professional development and district/school supports, and stability of the community and school), identifying the effects of the mentoring program on the principals' growth and performance as an instructional leader is just too complex to determine. A constructivist viewpoint would be satisfied with that position, and it is my understanding that the evaluators on my panel would have been satisfied with that as well. I came away from my conversations feeling like they were realistic and pragmatic in the expectations for the expected outcomes of our programs. Neither group—evaluators or program administrators—gave me the impression that they expected to see evidence of long-term outcomes that would extend out to improved student learning. The evaluators on the panel made the point that if an outcome is included in the program theory, then it must be tested. If stakeholders are expecting a long causal chain that extends out to long-term outcomes, then it is the evaluator's duty to warn the stakeholders of the difficulty of providing evidence for such long term outcomes. Mentoring program administrators doubted that an evaluation of a mentoring program could get much beyond identifying effects on leadership skills much less effects on schools and students.

With my constructivist paradigm in mind for this evaluation, particularly Dahler-Larson's (2001) call for describing the interdependence of the program and its context, I still managed to fail at effectively bringing in the contextual components of the program. I was criticized roundly for missing key components of the setting and program participants. In several of my journal entries, I felt the tension between constructivism and PTE. Constructivism celebrates complexity and multiplicity, while PTE strives for reductionism and parsimony. PTE looks for the underlying assumptions and theories that guide program implementation which is a good fit for a constructivist way of examining programs. However, because outcomes were highlighted on that

initial a priori program theory of the mentoring program, the pull to struggle and identify the outcomes of the program overcame the work to identify the contextual features and program processes.

Theme 2: Representing the complexity of programs. A tension that I saw repeatedly in my own evaluation as well as in the others was the tension between a graphical form of the program theory and honoring the rich description of the program presented by program stakeholders. Presenting a program theory, particularly in the graphic form that distilled the understanding about the program into a user-friendly visual can be very helpful in understanding how and why a program works. An analogy that would best described how I felt the program theory model guided me in my evaluation, as well as understand the programs in the comparison sample evaluations, was the program theory model as a GPS locator. With the pure constructivist evaluation of the gifted program (Ford, 1995), it was easy to get lost in the narrative. Thus, adding in the PTE elements to my mentoring program evaluation brought an advanced organizer into the report to help digest and understand the complex data produced during an evaluation. However, with those benefits in mind, representing a program—its context, processes, and outcomes—in a graphical program theory form was more difficult than expected, particularly if an evaluator does choose to identify and highlight the effects of the motivations and agendas of program stakeholders. For instance, in a graphical representation of a program, it was unclear of how to represent the values, beliefs and motivations of program stakeholders. In the evaluations in the comparison sample, I saw the assumptions that underlie a program theory listed in a text box above or below the logic model or program theory, or I have seen context encircling the entire program theory figure. Neither of those options was a true reflection of the impact that those values, beliefs, and motivations have on how stakeholders experience the program or the

program's effectiveness. In addition, because these contextual features do not figure prominently or intentionally in a program theory or logic model, I believe that is why they were often ignored or omitted from the program description.

Theme 3: Clarifying criteria of merit or worth. All three of my data methods found that clearly identifying criteria for judging program merit or worth was problematic in that clear criteria were not identified but often assumed. Both PTE and constructivist principles advise developing criteria out of the program (e.g., program goals and purpose). However, in mine and in the comparison samples, the evaluators did not clearly define these criteria for the reader. The evaluators on my metaevaluation panel who have written and worked primarily in the program theory approach were most concerned with clear measures of causation of effectiveness. In other words, implementing a comparative design to the study would have provided positive (or negative) evidence of whether the program met its goals and was thus, worthwhile, and it may have eliminated alternative explanations for the effects. These program theory and realist evaluators also were concerned with the potential bias of program stakeholders' self-reported data. They said that those stakeholders who liked the program may have shaped their responses to reflect more positively on the effects of the program than what may have been uncovered through more objective data such as teacher reports of effects on protégés' leadership behaviors in the school.

On the contrary, the mentoring program administrators wanted more in-depth descriptions of the foci and discussions that took place in mentoring programs to be able to determine if the program was high quality. For instance, the program administrators would have wanted to know what leadership skills the protégés and mentors were focusing on in the mentoring sessions as a determination of whether they believed the program was worthwhile.

Lastly, the constructivist evaluator on the panel was more concerned with the lived experiences of the stakeholders as evidence of merit and worth. As an example, this panelist asked the question, “If principals become more confident and doing positive things like praising teachers but there is no rise in student achievement, is it still a worthwhile program?” The protégés who participated in this study, through the survey and in the case study, said that they would recommend the mentoring program to their peers. From this constructivist standpoint, if we were to take the recommendations of the participants as evidence of its worth, then the answer would be yes. If we were to take another perspective from the standpoint of wanting evidence of outcomes, then the answer would be maybe, but more evidence is needed.

In the following chapter, I discuss my lessons learned from this metaevaluation based on these findings as well as the implications for the evaluation field.

Chapter 6

Lessons Learned and Implications

This dissertation constitutes a reflective and critical study of a constructivist program theory evaluation (C-PTE) approach used in the evaluation of a school district's new principal mentoring program. I chose to integrate these two evaluation theories for the following reasons. First, the constructivist paradigm oriented me toward the beliefs, assumptions, and values of stakeholders and this helped me understand how the program was implemented and the value district personnel placed on the program. Mentoring programs are reliant on the quality of the experiences and relationship between mentor and protégé, and the constructivist qualitative framework allowed me to explore these relationships to understand the relational and contextual dimensions of program quality. From what I considered about the relational aspects of mentoring programs and the complexity of educational leadership and schooling in general, I believed that the constructivist elements would elevate and honor the in-depth descriptions and lived experiences of program stakeholders as well as the complexity of programs. For instance, schools live at the crossroads of many contextual elements of the communities and families they serve (e.g., educational level of families, levels of economic stability). They are guided by the vision of education set by the school board and central administration. Within the schools, there are often complex professional and interpersonal dynamics between the school leadership and teachers that can affect relationships, and too often affect how well the school principal is able to promote a vision of learning for the school.

Second, the PTE approach focused the evaluation on uncovering the underlying theories of the mentoring program held by various stakeholders in terms of stating the problem the mentoring program was trying to solve, describing how the program was implemented, and

identifying the effects the program was having on principals, their schools, their students, and the district. The program theory elements of the evaluation, I thought, would help me organize the complexity of this mentoring program and how it was working in the district. The sections below highlight lessons I learned from this study and implications for the field of evaluation. The following section is a reflection of what I learned as a result of this metaevaluation: what I would do the same and what I would do differently..

Developing the Program Theory

This first set of lessons learned revolves around the evaluation process of developing and testing the program theory. There are two areas that were most salient in the metaevaluation: 1) attending to the contextual features surrounding the program and incorporating them in the program theory, and 2) identifying appropriate program outcomes and determining causality.

Returning to the constructivist critique of PTE. A constructivist critique that I cited in the literature review focused on the incorporation of context in PTE. Dahler-Larson's (2001) critique of PTE was that this approach did not attend to contextual features as much as it should, particularly the interdependence of a program and its context. In his conception of tragic, magic and competing programs, Dahler-Larson (2001) discussed the interactions of contextual features that may either facilitate a program's success or facilitate its demise. In the examples he gave, the contextual features were the political climate and social perceptions of the program stakeholders that moderated or inhibited program outcomes.

According to Greene (2005), context can be defined as "the setting in which the evaluand and the evaluation are situated. Context is the site, location, environment, or milieu for a given evaluand" (p. 83). Greene continued on to describe the multiple dimensions of the evaluand's context that include the demographic characteristics of the setting, economics and available

resources in the physical setting, institutional and organizational climate, the interpersonal interactions and relational norms, and political forces or influences (p. 83). In a special issue of the 2012 *New Directions for Evaluation*, evaluators wrote about the need for a greater emphasis on context in program evaluations and called for evaluators to become more context sensitive in their practice (Fitzpatrick, 2012; Rog, 2012). Citing Greene's (2005) definition of context and expanding on it further, Rog (2012) identified five areas of context to which evaluators should attend to build a framework of context for evaluation: context of the problem or phenomenon, context of the intervention, the broader environment or setting of the intervention, the parameters of the evaluation, and the decision-making context. Within each of these areas, she listed seven dimensions: physical, organizational, social, cultural, tradition, political, and historical (p. 27). Taking this idea of context sensitivity another step further, Conner, Fitzpatrick, & Rog (2012) used Rog's framework to develop a system of context assessment. These evaluators described an assessment system that incorporates context into the evaluation at three different phases: 1) assessing context during the evaluation planning stage, 2) monitoring for changes in the context during evaluation implementation, and 3) assessing the context associated with the use of the evaluation and surrounding the decision-making process and actors (Connor, et al., 2012).

Using this context assessment as a frame for analyzing the findings of the metaevaluation to extrapolate lessons, I returned to the claims both in my reflection journal and the responses of the metaevaluation panel that were related to (a) the difficulty of incorporating context into the program theory, and (b) the extent to which I offered a broad enough perspective in identifying relevant contextual features that may have impacted the program's implementation and outcomes. This breakdown, I believe, started with poor context assessment during the evaluation planning stage.

During the evaluation planning stage, I typically begin the inquiry process with a literature review and review of prior evaluations and research on programs. To organize the information I am learning about a program, developing crosswalks and logic models is a way that helps me process and integrate my learning. Even though the a priori development and testing of the program theory was confusing for many of my panelists in that it was difficult to determine when development ended and testing began, I valued having this theory at the outset of the evaluation to help me develop the evaluation questions and data collection instruments. However, upon reflecting on the evaluation in my journal and from the metaevaluation panel responses, I am wary of the bias that the literature review and a priori development of a program theory without stakeholder input may bring into an evaluation. I need to be more conscious of the constructivist elements in my evaluation. The a priori program theory may have constricted my vision, which may be a large reason for why the stakeholder theories converged so easily. At this writing a few years away from the evaluation, it remains a question as to why I did not ask the basic question from each stakeholder of how they defined school leadership, particularly the mentors and protégés. Asking this basic question of these two primary program stakeholders may have better facilitated the discussion about how the mentoring program was helping the protégés meet this vision of the principalship. It also would have been interesting to juxtapose their visions of the principalship versus the district administrators' visions.

Furthermore, with a deeper analysis of the context with a broader set of stakeholders in the evaluation planning stage, it may have become more evident that the evaluation needed to focus on developing the theory and that the program was not ready to test the theory to evaluate its impact on the protégés, their schools, or their students. Early on in the evaluation interviews with the program administrators, it became evident that the program had been hastily

incorporated into the district's plan for improving the quality of their principals as part of an externally funded grant. At one point when asked, the program administrators were not clear about what the desired goals and outcomes of the mentoring program should be beyond providing support to the principals and improving their practice. However, the indicators, or criteria, for what improved practice should look like had not been explicitly identified or communicated to program stakeholders. Therefore, I should have taken this revelation as a cue to draw on the strengths of constructivist evaluation to go deeper into the constructs that define effective school leadership that the district administrators, as well as school board members, principals, mentors, teachers, students, and community members each had in mind. Going back to Dahler-Larsen's (2001) point that these social constructs may either facilitate the program's success or failure, competing constructs of school leadership may have caused this program to struggle if these groups had different beliefs and expectations for what effective school principals should do and how they should act when leading their schools. If these different stakeholders had competing constructs (e.g., a management orientation versus an instructional leadership orientation), then the probability for the program's success would be minimized. For example, if the mentors and protégés worked from an instructional leadership orientation and focused on improving the protégé's capacity to improve aspects of the curriculum and teaching practices in the school, but the school board members and district administrators were more concerned with the protégé's ability to manage school schedules and budgets, then the mentoring program may be seen as a failure in the eyes of the district administrators and school board. In contrast, if all the program stakeholders are operating under the same set of constructs for school leadership, the mentoring program might have a better chance for success.

Furthermore, I would also choose to focus on studying the fidelity of implementation of the mentoring program. First, I would work with the program administrators to define the indicators of fidelity of implementation, and then from a practical standpoint, choose a couple protégés to study what fidelity of implementation looks like in the field. Conducting this type of data collection and analysis would have enabled me to better identify program practices that the program administrators on my metaevaluation panel would have appreciated in opening the “black box” of those mentoring sessions. It also would have better enabled me to identify and explore the causal linkage between the program practices, mediators, and short-, intermediate, and long-term outcomes.

On a political and economic level, the metaevaluation panelists were justified in critiquing my narrow contextual lens in which I did not include school board members or community members in the evaluation planning or data collection process. In 2011, the Illinois State Board of Education (ISBE), due to economic hardships in Illinois, did not allocate funds to the Illinois New Principal Mentoring Program, which meant that beginning in FY12, districts would no longer receive funding from ISBE to pay for their new principals’ mentoring. Therefore, the new principal mentoring mandate no longer applied, and mentoring was voluntary. Some school districts, like Chicago Public Schools and other districts in the northern parts of Illinois, continued to offer mentoring for new principals. However, with a turnover in district superintendent and no appropriation from the ISBE, the Clearlake School District discontinued the new principal mentoring program. This discontinuation was a signal that in the district some stakeholders did not value the program. A better contextual assessment in the evaluation may have predicted its demise if certain political and economic factors would change and provided better information for stakeholders and their decision-making process.

Measuring outcomes and evaluating causality. Directly related to my discussion about context, the difficulty in identifying outcomes and showing causality was a result of not spending enough time in assessing the contextual features of the program. I felt that the program theory pulled me toward looking for outcomes to the extent that if outcomes could not be verified, then the evaluation would be a failure. In my interviews with my metaevaluation panel, I was continually surprised (and pleased) about the value that the panelists placed on the contextual and descriptive elements of the evaluation, even though they were critical of my efforts in this respect. As the only evaluator on the project with very little funding, this search for outcomes took up time and attention that would have been better served in going deeper in the contextual and descriptive analysis of the program's functioning. In my reflective journal and my metaevaluation panel, there was great concern that I did not go deep enough into what the program looked like. For instance, on what instructional leadership practices were mentors and protégés focusing during their sessions? What happened in these sessions? Were they focused on personal issues, leadership issues, both, neither? Did the mentor and protégés share the same goal for the program as the district administrators? Did the school board members support the mentoring program? (Did they even know about the program?) I think that the value placed on the contextual and descriptive piece justifies the importance of allocating evaluation resources to spending time in the field collecting these data so that a) the evaluator knows whether it is feasible to collect outcomes data, and b) what outcomes data are appropriate at this time and for this program.

With that last statement in mind, I was appropriate in abandoning the collection of student outcomes data. It was too soon for many of the participating protégés in my evaluation as they had only been in their positions for two or three years, and general student test scores may

not have served an outcomes or ‘effectiveness’ purpose in my evaluation. It would have been better practice to work at a deeper level with the mentors and protégés to uncover initiatives that the protégés believed the mentoring program helped them implement in the schools and then collect the data that may have substantiated those claims. Working at this deeper level would have also helped me target the types of questions to ask the teachers in the building in terms of what differences the principal was making in the school. It would be erroneous to believe that a principal in his/her second or third year on the job could make wholesale changes to the working conditions in the school. However, there may have been a few areas where the principal was specifically targeting that would have been interesting to observe and question.

If I or other evaluators were to conduct an evaluation of new principal mentoring programs in Illinois today, we would have some better data sources in our toolbox that may help target and measure changes in program outcomes. First, in 2009, the Illinois General Assembly (ILGA) passed the P-20 Longitudinal Education Data System Act (P. A. 096-0107) that directed the Illinois state education agencies to work together to develop an aligned system of data that would allow policymakers and researchers to study and evaluate the characteristics of the State’s education programs, education workforce, and students. At this writing the longitudinal data system is still under development, but once it is finished, the variables included in the system may provide a more efficient and quality avenue for drawing down data that could be used to describe certain contextual factors and program outcomes (e.g., demographics, economic factors, student assessment data). In addition, in 2010 the Illinois General Assembly passed the Performance Evaluation Reform Act (P. A. 096-0861) that reformed the State’s principal and teacher evaluation systems. This act included a provision that the Performance Evaluation Advisory Committee should develop the guidelines and standards for principal and teacher

evaluation to which all school districts have to comply. For principals, this includes a set of school leadership standards by which principals are evaluated according to four performance levels (unsatisfactory, basic, proficient, and excellent) (ILGA, 2010). By setting these standards, the ILGA set a common vision of school leadership that could be used in a comparative analysis of principal performance in districts across Illinois. In addition, this Act mandated that an annual survey of learning conditions be implemented in every district, and the ISBE now mandates that every district for every school in their district distribute the 5 Essentials Survey that collects learning condition data from teachers, students, and parents (ISBE, n.d.). These data also hold promise for being able to measure the effects that principals are having on improving these conditions in their schools, and thus, the impact that programs like mentoring, have on the principals' capacity to make those improvements.

Benefits of the Metaevaluation Panel

Through this reflective metaevaluation study, I quickly came to realize the benefits of convening a metaevaluation panel, not just as a quality control mechanism, but also for bringing in different viewpoints into the evaluation. For instance, I thought it was an interesting juxtaposition that while both evaluators and program administrators on the panel said that they wanted to see more depth in the descriptions of program processes, the evaluators were much more concerned than program administrators with identifying causal mechanisms and valid findings (i.e., cross-case comparisons and comparison group analyses). The evaluators still saw the value of context as a way to justify inferences of outcomes. However, program administrators wanted depth as a way of learning new program strategies to bring back to their programs, or justifying their program's practices by rejecting the ones deemed ineffective in the evaluation. The program administrators were less critical of the outcomes data because as part of their

experience as school and district leaders, and now mentoring program administrators, they knew the difficulty of measuring outcomes and attributing cause.

Having the mix of evaluators and program administrators of similar programs offered a well-rounded set of perspectives that offered crucial feedback on my evaluation processes and findings. It would be valuable to convene such a panel at different points of an evaluation: during the evaluation design, in the middle of data collection to discuss the evaluation process and preliminary findings, and after the writing of the final report.

Criteria for Judging Constructivist Program Theory Evaluations

After reflecting on this study, there were certain criteria that were more helpful in the metaevaluation of the C-PTE of the new principal mentoring program than others. I will begin by describing those criteria that I would recommend use in future evaluations using a C-PTE evaluation model. In addition, the Joint Committee on Standards for Educational Evaluation updated the standards between the time of my metaevaluation and this writing (2011). So, where appropriate, I will reference criteria from the 3rd edition of these standards where I believe they should be incorporated.

Context program interdependence. The first criterion that I would recommend comes from the constructivist field and is the criterion that I labeled context-program interdependence. This was a criterion that I created in response to Dahler-Larsen's (2001) article that recommended a two-way examination of the relationship between the context and program. So, this criterion would suggest that the evaluator should investigate where the program sits within the context and how the context affects program operations and success, as well as how the program changes the context, which could be a change in attitudes or beliefs, a change in policies or practices, or an improvement (or decline) in a measure of performance (e.g., improved school

climate). With regards to context, the 3rd edition of the Joint Committee on Standards (2011) included three standards, Explicit Values, Explicit Program and Context Descriptions, Contextual Viability (p. 37, p. 185 & p. 93, respectively), that look to be promising in holding evaluators accountable for including context into program evaluations. In order to conduct evaluations that meet this criterion to an exemplary level, the evaluation field would do well to incorporate Rog's (2012) framework and Connor and his colleague's (2012) context assessment framework for further investigation into expanding on the Joint Committee's descriptions of context and including Dahler-Larsen's call for exploring the interdependency of contexts and programs.

Principled discovery and competitive elaboration. In the literature (Julnes & Mark, 1998), recommended that evaluators should use either principled discovery or competitive elaboration for purposes of assessing how well the evaluator was systematic in their data collection and analysis to elicit valid and reliable findings. Principled discovery was described as a process of inductive and deductive inquiry that allows for exploratory data collection and analysis that is bounded by the parameters of deductive data analysis (Julnes & Mark, 1998). Competitive elaboration was described as the evaluator's ability through data collection and analysis to identify and account for alternative explanations for inferences made from the data about program practices and outcomes (Julnes & Mark, 1998). Furthermore, Julnes and Mark (1998) presented principled discovery as more appropriate if the program is relatively new or there has not been much prior evaluation or research on the program. Competitive elaboration, on the other hand, was recommended if there has been adequate study of the program, and then the focus would be on alternative explanations that may have been previously identified in prior evaluations through principled discovery. I included both and defined principled discovery as the

systematic investigation of the program processes and outcomes and competitive elaboration as ruling out alternative explanations. The metaevaluation found that my evaluation met the criteria of principled discovery to some extent, primarily in exploring program practices. But, my evaluation was not rated as highly in these two criteria when it came to confirming causality between the program practices and outcomes and in ruling out alternative explanations for the evaluation findings.

In the PTE and realist evaluation literature, it would be the goal to rule out alternative explanations so that the evaluator would have a reasonable expectation that the findings could be attributed to the program. I felt and still feel that from the constructivist viewpoint, alternative explanations should be explored as a way of having a deeper understanding of the program. However, in relation to the complexity of the mentoring program that was the evaluand, I am not sure that the alternative explanations could ever be explained away. In order to explain the explanations away, at the very least, an evaluation would have to be focused on very specific leadership knowledge or behaviors that the mentoring program was seeking to address, and identify sensitive measures that would assess changes in those knowledge or behaviors to determine causality. The authors of the 3rd edition of the Joint Committee on Standards for Educational Evaluation (2011) wrote a standard they called Reliable Information that calls for evaluators to examine for sources of error in evaluation findings (e.g., alternative explanations that cannot be identified or alternative explanations that cannot be accounted for (i.e., explained away) with the data) (p. 179). . The writers of this standard hold evaluators accountable to be systematic in developing data sources that increase the reliability of findings and decrease error which is similar to the principled discovery and competitive elaboration criteria I used for this metaevaluation.

Resonance and transferability. These two criteria assessed the extent to which the program evaluation produces a rich description of the program that elicits fuller understandings of the program for the reader, and whether the understanding of this program can translate to other similar programs. These criteria fit the utility concerns of both constructivism and program theory evaluation well, and metaevaluation panelists had no trouble responding to these criterion and rated my evaluation highly on both criteria. The 3rd edition of the Joint Committee on Standards for Educational Evaluation (2011) included several standards that address the resonance and transferability criterion. These are the standards of Relevant Information, Meaningful Processes, and Products (p. 45 & 51). Both of these standards are associated with producing evaluation information that is useful for the program stakeholders and intended users.

Fairness. Finally, the fairness criteria should be an essential criterion regardless of what type of evaluation approach is chosen to evaluate a program. Both constructivist and program theory approaches can be well suited for ensuring that all voices are included in the evaluation if the evaluator attends to this criterion with due diligence and close examination. As mentioned earlier, my evaluation came to easy convergence on the program theory where I expected multiple program theories. So, it begged to question whether the fairness criterion was addressed well. The 3rd edition of the Joint Committee on Standards for Educational Evaluation (2011) included a standard of Clarity and Fairness (p. 131) that supports the meeting of this criterion in every evaluation. Furthermore, the Joint Committee (2011) also included a standard of Contextual Viability (p. 93) that holds evaluators accountable for assessing the political and cultural interests of program stakeholders to ensure that all relevant stakeholders' perspectives and needs are addressed during the evaluation and no one stakeholder group's needs are integrated into an evaluation over others'.

Credibility. The criterion that I tended not to find as useful for this evaluation is the credibility criterion. I found that it encompassed many of the others (e.g., fairness, principled discovery, and competitive elaboration). Therefore, it seemed redundant, and it was difficult to engage this criterion in my reflection journal as well as with my metaevaluation panel. The analysis of the panelists' comments on the credibility criterion compared to the comments on the fairness, principled discovery, and competitive elaboration criteria showed a pattern of nearly identical responses. Furthermore, information scope and selection did not elicit particularly interesting responses from the mentoring program administrators on my metaevaluation panel and was not an issue that I discussed in my reflection journal. I also did not use this criterion in my comparative analysis because I did not feel like I knew enough about the programs and fields in which my comparison sample resided to judge them on this criterion. Upon reading the 3rd edition of the Joint Committee on Standards for Educational Evaluation (2011), I noticed that credibility descriptions tend to be disbursed throughout the standards in different attributes and standards throughout the text, while utility is an attribute that is given its own set of standards. Therefore, it would appear that the Joint Committee also felt, like I, that credibility is a dimension that cuts across many different criterion of quality evaluations.

The next section highlights my thoughts about how this study informs where the evaluation field may learn from this study.

Implications for the Evaluation Field

Implications for Research and Practice

Rog (2012) continued the call for program evaluators to reflect on and research evaluation practice to better understand and refine our program evaluation theories and approaches. Evaluators who want to pursue the blending of two evaluation models, particularly

in this case constructivism and PTE, need to be cognizant of the tensions and consonances that exist between components of the two models, and have a specific plan to monitor and be intentional in the mixing of the two models to ensure that the resulting evaluation engages sufficiently with the characteristics of the context and meets the information needs of stakeholders. With that being said, I would advise the evaluation field to integrate evaluation approaches as a way to study and test our evaluation theories. In the tensions between two approaches, strengths and weaknesses of evaluation models tend to become more obvious. Tensions tend to arise in several areas where evaluators have to make decisions about what type of questions to ask, methods to use, types of data to collect and analyses to conduct, and criteria for judging program merit and worth as well as for judging the credibility and utility of the evaluation. Some of these areas of tension that may arise as evaluators ask their design questions may include decisions around: context data and program data, process data and outcomes data, explanation and causation, responsive methods and emerging data, fixed (i.e., predetermined) evaluation methods and data, role of stakeholders in the evaluation, role of the evaluators, criteria for judging merit and worth of the program, criteria for judging credibility and utility of the evaluation. When integrating two or more evaluation theories, it may not always be a deliberation of one element versus another (e.g., explanation versus causation), but it may instead be how does the evaluator do both, should s/he do both.

Implications for Evaluator Training

Similarly, I believe it would be of value for graduate programs that train future evaluators to include learning experiences in courses or practical experience to integrate models of evaluation theory as a way of learning about the differences in these theories. In designing an evaluation by integrating models, evaluators in training will inevitably produce tensions when

making choices about evaluation questions, methods, instrument development, analyses and making inferences. Through this experience, future evaluators can study what happens when you integrate evaluation models—what gets compromised, what gets enhanced—which may also be a useful way to teach future evaluators about evaluation theory as well as the development, implementation, and reflection cycle that all professionals should engage in order to advance their skills.

In Conclusion

Conducting an integrated evaluation approach through the constructivist program theory evaluation of the Clearlake School District new principal mentoring program has been a challenging exercise in exposing and working within the tensions created by blending the two evaluation models. Blending evaluation theories, particularly theories that have serious dichotomies, is strenuous and intellectually challenging in making clear choices in the evaluation design and implementation phases to ensure credible and useful evaluation findings. When integrating theories, evaluators should be cognizant that the strengths of one evaluation theory may dominate over the elements of the other evaluation approach that may disenfranchise the strengths of the other evaluation approach and enhance its weaknesses. One way to avoid this pitfall is to convene a metaevaluation panel that includes a breadth of evaluation theoretical perspectives including those who typically practice in the evaluation approaches that were integrated. The panel may also include other evaluators with another theoretical or pragmatic perspective to help the evaluator think beyond the theoretical perspectives incorporated in the evaluation. The panel should also include practitioners and stakeholders who work or reside in similar programs to bring yet another perspective to the evaluation. Finally, the evaluator should

be very deliberate in choosing criteria to assess the evaluation's quality and utility that both fit and challenge the evaluation theories used to conduct the evaluation.

I hope that the evaluation field continues to support metaevaluation studies to assess the value of our evaluation theories as they are practiced. While it can be uncomfortable to offer up one's evaluation practice to a panel of peers, the lessons learned are invaluable. In my own evaluation practice as I have seen program theory evaluation and logic models grow in demand, I will continue to advocate for the contextual analyses and in-depth qualitative data collection of stakeholders' experiences with the programs so that inferences of program effectiveness can be more credibly attributed to the essential elements of programs.

References

- Adler, M. A. (2002). The utility of modeling in evaluation planning: the case of the coordination of domestic violence services in Maryland. *Evaluation and Program Planning, 25*, 203-213. doi:10.1016/S0149-7189(02)00016-2
- Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation, 33*, 5-18. doi: 10.1002/ev.1443
- Bickman, L. (1989). Barriers to the use of program theory. *Evaluation and Program Planning 12*, 387-390. doi: 10.1016/0149-7189(89)90056-6
- Bickman, L. (1996). The application of program theory to the evaluation of a managed mental health care system. *Evaluation and Program Planning, 19*, 111-119. doi: 10.1016/0149-7189(96)00002-X
- Birckmayer, J. D. & Weiss, C. H. (2000). Theory-based evaluation in practice. *Evaluation Review, 24*, 407-431. doi: 10.1177/0193841X0002400404
- Bowen, K. A. (1999). *Evaluation of a teen pregnancy and parenting program: Integrating an interpretivist, theory-oriented approach* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (AAT 9923763).
- Browne-Ferrigno, T. & Muth, R. (2004). Leadership mentoring in clinical practice: Role socialization, professional development, and capacity building. *Educational Administration Quarterly, 40*, 468-494. doi:10.1177/0013161X03255561
- Bloom, G., Castagna, C., Moir, E., & Warren, B. (2005). *Blended coaching: Skills and strategies to support principal development*. Thousand Oaks, CA: Sage.

- Brandon, P. R., & Fukunaga, L. L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35*, 26-44. doi:10.1177/1098214013503699
- Calvin, L., Streeter, C. F., Kim, J. S., & Tripodi, S. J. (2011). Concept mapping: An approach for evaluating a public alternative school program. *Children & Schools, 33*, 197-214. doi:10.1093/cs/33.4.197
- Cash, T. & Cloud-Duttweiler, P. (2005). *Planning, Collaboration, and Implementation Strategies for Truancy Programs*. Clemson, SC: National Dropout Prevention Center/Network.
- Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509-535). Thousand Oaks, CA: Sage.
- Charmaz, K. (2005). Grounded theory in the 21st century: A qualitative method for advancing social justice research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (3rd ed., pp. 507-535). Thousand Oaks, CA: Sage.
- Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: Sage.
- Chen, H. T. (1989). The conceptual framework of the theory-driven perspective. *Evaluation and Program Planning, 12*, 391-396. doi:10.1016/0149-7189(89)90057-8
- Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.
- Chen, H. T. & Rossi, P. H. (1992) (Eds.). *Using theory to improve program and policy evaluations*. Westport, CT: Greenwood Press.

- Christie, C. A. & Alkin, M. C. (2003). The user-oriented evaluator's role in formulating a program theory: Using a theory-driven approach. *American Journal of Evaluation*, 24, 373-385. doi:10.1177/109821400302400306
- Coffey, A. & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. Thousand Oaks, CA: Sage.
- Conrad, K. J. & Miller, T. Q. (1987). Measuring and testing program philosophy. *New Directions for Program Evaluation*, 33, 19-42. doi:10.1002/ev.1444
- Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, 24, 119-128. doi:10.1016/S0149-7189(01)00003-9
- Crane, B. (2000). *Building a theory of change and a logic model for an empowerment-based family support training and credentialing program* (Doctoral dissertation). Retrieved from ProQuest Dissertation and Theses Database. (AAT9967470).
- Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.
- Dahler-Larson, P. (2001). From programme theory to constructivism: On magic, tragic and competing programmes. *Evaluation*, 7, 331-349. doi:10.1177/13563890122209711
- Daresh, J. C. (2004). Mentoring school leaders: Professional promise or predictable problems? *Educational Administration Quarterly*, 40, 495-517. doi:10.1177/0013161X04267114
- Daresh, J. C., & Playko, M. A. (1993). Benefits of a mentoring program for aspiring administrators. Paper presented at the meeting of the American Association of School Administrators, Orlando, FL. (Document Reproduction Service No. ED354603)

- Davidson, E. J. (2006). The “bagging” of theory-based evaluation. *Journal of MultiDisciplinary Evaluation*, 4, iii-xiii.
- Donaldson, S. I. (2003). Theory-driven program evaluation in the new millennium. In S.I. Donaldson and M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the New Millennium* (pp. 109-144). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Donaldson, S. I. & Gooler, L.E. (2003). Theory-driven evaluation in action: Lessons from a \$20 million statewide Work and Health initiative. *Evaluation and Program Planning*, 26, 355-366. doi:10.1016/S0149-7189(03)00052-1
- Donaldson, S. I. & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In F. Shaw, J. C. Greene, & M. M. Mark (eds.), *The SAGE Handbook of Evaluation*. London: Sage.
- Ehrich, L. C., Hansford, B., & Ehrich, L. (2005) Formal mentoring programs in education and other professions: A review of the literature. *Educational Administration Quarterly*, 40, 518-540. doi:10.1177/0013161X04267118
- Ford, R. (1995). *A non-traditional evaluation of the Woodbury Public School Elementary Enrichment Program* (Doctoral dissertation). Retrieved from ProQuest Dissertation and Theses Database. (AAT 9527469).
- Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*. San Francisco, CA: Wiley & Sons, Inc.
- Friedman, V. J. (2001). Designed blindness: An action science perspective on program theory evaluation. *American Journal of Evaluation*, 22, 161-181.
doi:10.1177/109821400102200203

- Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Wiley & Sons, Inc.
- Enomoto, E. K., Gardiner, M. E., & Grogan, M. (2000). Notes to Athena: Mentoring relationships for women of color. *Urban Education, 35*, 567-583.
doi:10.1177/0042085900355007
- Geertz, C. (1973). *The interpretation of cultures*. NY: Basic Books.
- Geismer, T. J., Morris, J. D., & Lieberman, M. G. (2000). Selecting mentors as principal interns. *Journal of School Leadership, 10*, 233-247.
- Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Green, B. L. & McAllister, C. (2002). Theory-based, participatory evaluation: A powerful tool for evaluating family support programs. *The Bulletin*. Retrieved from <http://www.npresearch.com/Files/SRCDZ3.pdf>
- Greene, J. C. (1993). The role of theory in qualitative program evaluation. In D.J. Flinders and G.E. Mills (Eds.), *Theory and concepts in qualitative research* (pp. 24-45). NY: Teachers College Press.
- Greene, J. C. (2005). Context. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.
- Guba, E. G. & Lincoln, Y. S. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.
- Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.
- Hallinger, P. & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of the empirical research, 1980-1995. *Educational Administration Quarterly, 32*, 5-44. doi:10.1177/0013161X96032001002

- Hansford, B. & Ehrich, L. C. (2005). The principalship: How significant is mentoring? *Journal of Educational Administration*, 44, 36-52. doi:10.1108/09578230610642647
- Hendricks, M., Plantz, M. C., & Pritchard, K. J. (2008) Measuring outcomes of United Way funded programs: Expectations and reality. *New Directions for Evaluation*, 119, 13-35. doi: 10.1002/ev.266
- Henry, G. T. & Mark, M. M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation*, 97, 69-80. doi:10.1002/ev.77
- Howe, K. R. (2004). A critique of experimentalism. *Qualitative Inquiry*, 10, 42-61. doi:10.1177/1077800403259491
- Illinois General Assembly (2006, July). Illinois New Principal Mentoring Program. (Illinois Public Act 094-1039). Springfield, IL: Author. Retrieved from <http://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=094-1039>
- Illinois General Assembly (2009, July). P-20 Longitudinal Education Data System Act. (Illinois Public Act 096-0107). Springfield, IL: Author. Retrieved from <http://www.ilga.gov/legislation/publicacts/96/096-0107.htm>
- Illinois General Assembly (2010, November). Performance Evaluation Reform Act: Laws of Illinois, 2010. (Illinois Public Act 096-0861). Springfield, IL: Author. Retrieved from <http://www.ilga.gov/legislation/publicacts/96/pdf/096-0861.pdf>
- Illinois State Action for Education Leadership Project. (2005). Unpublished raw data. Illinois State University: Normal, IL.

- Illinois State Board of Education. (no date). 5 Essentials Survey. Retrieved from <http://www.isbe.net/5essentials/default.htm>
- Jasper, M. A. (2005). Using reflective writing within research. *Journal of Research in Nursing, 10*, 247-260. doi:10.1177/174498710501000303
- Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*, 377-410. doi: 10.1177/1098214009341660
- Joint Committee on the Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks, CA: Sage.
- Joint Committee on the Standards for Educational Evaluation. (2010). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.
- Julnes, G. & Mark, M. M. (1998). Evaluation as sensemaking: Knowledge construction in a realist world. *New Directions for Evaluation, 78*, 33-52. doi:10.1002/ev.1099
- Kaplan, S. A. & Garrett, K.E. (2005). The use of logic models by community-based initiatives. *Evaluation and Program Planning, 28*, 167-172.
- Kellogg Foundation, The. (2004). *Logic model development guide*. Battle Creek, MI: The Kellogg Foundation. Retrieved from <http://www.wkcf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide>
- Kent, D. R., Donaldson, S. I., Wyrick, P. A., & Smith, P. J. (2000). Evaluating criminal justice programs designed to reduce crime by targeting repeat gang offenders. *Evaluation and Program Planning, 23*, 115-124. doi:10.1016/S0149-7189(99)00039-7

- King, J. A. (2003). The challenge of studying evaluation theory. *New Directions for Evaluation*, 97, 57-68. doi:10.1002/ev.76
- Kvale, S. (1996). *Inter views: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.
- Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. NY: The Wallace Foundation.
- Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions for Program Evaluation*, 57, 5-38. doi:10.1002/ev.1637
- Marchal, B., Dedzo, B., & Kegels, G. (2010). A realist evaluation of the management of a well-performing regional hospital in Ghana. *BMC Health Services Research*, 10(24), Retrieved from <http://www.biomedcentral.com/1472-6963/10/24>
- Mark, M. M., Henry, G. T., & Julnes, G. (1998). A realist theory of evaluation practice. *New Directions for Evaluation*, 78, 3-32. doi:10.1002/ev.1098
- Mark, M. M., Hofmann, D. A., & Reichardt, C. S. (1992). Testing theories in theory-driven evaluations: (Tests of) moderation in all things. In H. T. Chen and P. H. Rossi (Eds.), *Using theory to improve program and policy evaluations* (pp. 71-84). NY: Greenwood Press.
- Maxwell, J. A. (1996). *Qualitative research design*. Thousand Oaks, CA: Sage.
- National Policy Board for Educational Administration. (2008). *Educational Leadership Policy Standards: ISLLC 2008*. Washington D.C.: Chief Council of State School Officers. Retrieved from http://npbea.org/wp-content/uploads/2012/06/Educational_Leadership_Policy_Standards_2008.pdf
- Pawson, R. & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

- Petrosino, A. (2000). Mediators and moderators in the evaluation of programs for children: current practice and agenda for improvement. *Evaluation Review*, 24, 47-72.
doi:10.1177/0193841X0002400102
- Plantz, M. C., Greenway, M. T., & Hendricks, M. (1997). Outcome measurement: Showing results in the nonprofit sector. *New Directions for Evaluation*, 75, 15-30. doi:
10.1002/ev.1077
- Playko, M. A. (1995). Mentoring for educational leaders: A practitioner's perspective. *Journal of Educational Administration*, 33, 84-92. doi:10.1108/09578239510098554
- Reynolds, A. J. (1998). Confirmatory program evaluation: A method for strengthening causal inference. *American Journal of Evaluation*, 2, 203-221.
doi:10.1177/109821409801900204
- Rich, R. A. & Jackson, S. H. (2005, May/June). Peer coaching: Principals learning from principals. *Principal*, 30-33. Retrieved from
<https://www.naesp.org/resources/2/Principal/2005/M-Jp30.pdf>
- Richardson, L. (2000). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 923-948). Thousand Oaks, CA: Sage
- Richardson, L. & St. Pierre, E. (2005). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (3rd ed., pp. 959-978). Thousand Oaks, CA: Sage.
- Rogers, P. J. (2000). Program theory: No whether programs work, but how they work. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.) (pp. 209-232). Boston: Kluwer Academic Publishers.

- Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsí, T. A. (2000). Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation*, 87, 5-13.
doi:10.1002/ev.1177:
- Rosas, S. R. (2005). Concept mapping as a technique for program theory development: An illustration using family support programs. *American Journal of Evaluation*, 26, 389-401.
doi:10.1177/1098214005278760
- Schwandt, T. A. (1993). Theory for the moral sciences: Crisis of identity and purpose. In D. J. Flinders & G. E. Mills (Eds.), *Theory and concepts in qualitative research* (pp. 5-23). NY: Teachers College Press.
- Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation*, 26, 95-105. doi:10.1177/1098214004273184
- Scriven, M. (1998). Minimalist theory: The least practice requires. *American Journal of Education*, 19, 57-70. doi:10.1177/109821409801900105
- Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation*, 19, 1-19. doi:10.1177/109821409801900102
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.
- Sheirer, M. A. (1987). Program theory and implementation theory: Implications for evaluators. *New Directions for Program Evaluation*, 33, 59-76. doi:10.1002/ev.1446
- Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *American Journal of Evaluation*, 14, 237-402.
doi:10.1177/109821409301400302

- Strauss, A. & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Stufflebeam, D. L. (2001a). Evaluation models. *New Directions for Evaluation*, 89, (complete issue). doi:10.1002/ev.3 and doi:10.1002/ev.4
- Stufflebeam, D. L. (2001b). The metaevaluation imperative. *American Journal of Evaluation*, 22, 183-209. doi:10.1177/109821400102200204
- Trochim, W. M. K. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning*, 12, 355-366.
- Trochim, W. M. K. & Cook, J. A. (1992). Pattern matching in theory-driven evaluation: A field example from psychiatric rehabilitation. In H. Chen, & P. H. Rossi (Eds.), *Using theory to improve program and policy evaluations* (pp. 49-69). NY: Greenwood.
- United Way of America. (1999). *Measuring program outcomes: A practical approach*. Alexandria, VA: United Way of America.
- Villani, S. (2006). *Mentoring and induction programs that support new principals*. Thousand Oaks, CA: Corwin Press.
- Wallace Foundation, The, (2007). *Getting principal mentoring right: Lessons from the field*. NY: The Wallace Foundation.
- Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.
- Weiss, C. H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review*, 21, 501-524. doi:10.1177/0193841X9702100405

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd Ed.). New Jersey: Prentice Hall.

Weiss, C.H. (2000). Which links in which theories shall we evaluate? *New Directions for Evaluation*, 87, 35-55. doi:10.1002/ev.1180

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation*, 22, 409-418. doi:10.1177/109821400102200319

Yampolskaya, S., Nesman, T. M., Hernandez, M., & Koch, D. (2004). Using concept mapping to develop a logic model and articulate a program theory: A case example. *American Journal of Evaluation*, 25, 191-207. doi: 10.1177/109821400402500204

Zellner, L. J., Skrla, L., & Erlandson, D. A. (2001). The art of mentoring principals.

Connections, 3. Retrieved from <http://www.nassp.org/portals/0/content/48348.pdf>

Appendix A

**Survey of Mentored Principals, Fall 2006
Illinois Mentoring Programs**

Demographic Information

Please answer the following:

1. Gender: Male Female

2. Ethnicity:

- White/Caucasian Black/African American
 Hispanic Asian American/Pacific Islander
 Native American Bi-racial/Multi-Ethnic

3. How many years have you worked in education? _____

4. What is your current title? _____

5. How many years have you worked in this position at this school? _____

6. In which coaching/mentoring/induction program did you participate?

- Program A⁶ Program C
 Program B Program D
 Clearlake School District

7. In what year(s) did you participate in this coaching/mentoring program?

- 2001-2002 2002-2003
 2003-2004 2004-2005
 2005-2006

8. Have you had other coaching/mentoring/induction experiences beyond the experiences in the program you identified in Question #6? If so, please indicate those that apply.

Informal mentor Internal district or school mentoring program Other

9. Please indicate the level and location of the school where you currently work.

Level

- Elementary
 Middle/Junior High
 High School
 Other

Location

- Chicago
 Chicago's collar counties (DuPage, Kane, Lake, McHenry, Will)

⁶ Pseudonyms replaced the names of the mentoring programs that were sampled for this survey to protect their identities.

- Small Urban (city with population over 25,000)
- Rural/Small Town (town with population under 25,000)

Mentoring Programs and Their Impact on Leadership Skills

For the purposes of this survey, mentoring program refers to a formal program (coaching, mentoring or mentoring/induction program) and specifically to the program that you selected in question #6. For the questions that follow, we are not asking you to reflect on your experiences in an informal mentoring relationship or another mentoring relationship provided by your school or district. **For the purposes of this survey, the term mentoring is used broadly for simplicity sake and encompasses coaching and mentoring/induction experiences in which you engaged in your coaching/mentoring/induction program.**

10. To begin, to what degree did the mentoring program impact your knowledge and skill in the following leadership behaviors? Please indicate the level of impact by circling the number that best describes the level of impact using this scale:

-2 -1 0 1 2

Negative impact Somewhat negative impact No impact Somewhat positive impact Positive impact

Responsibilities/Competencies of a School Leader	Degree of Impact
Establish a school mission with clear goals on improving student achievement and keeping those goals as the primary focus in the school	N/A -2 -1 0 1 2
Foster and communicate a school culture of shared beliefs and ideals about schooling that promotes a sense of community and cooperation and guides school operations	N/A -2 -1 0 1 2
Establish a set of standard operating procedures and routines to efficiently manage my school for a safe and orderly environment	N/A -2 -1 0 1 2
Develop and enforce discipline procedures that protect teachers from issues and influences that would detract from their teaching time or focus	N/A -2 -1 0 1 2
Adopt a paradigm of distributed leadership by involving teachers in the design and implementation of important decisions and policies	N/A -2 -1 0 1 2
Adapt leadership behaviors to the needs of the current situation and feel comfortable with dissent in regards to my actions	N/A -2 -1 0 1 2
Become situationally aware of the details and undercurrents in the running of the school and use this information to address current and potential problems	N/A -2 -1 0 1 2
Provide professional development and other learning opportunities to faculty and staff so they learn the most current theories and practices and make the discussion of these a regular aspect of the school's culture	N/A -2 -1 0 1 2
Monitor the effectiveness of school practices and evaluate their impact on student learning using multiple points of data	N/A -2 -1 0 1 2
Create opportunities to remain visible in the school and establish strong lines of communication with teachers and students	N/A -2 -1 0 1 2
Recognize and celebrate school accomplishments, using contingent rewards when appropriate, and acknowledge failures, affirming faculty, staff, and student success and encouraging them to make improvements in their teaching and learning	N/A -2 -1 0 1 2
Serve as a change agent to actively challenge the status quo to inspire and lead new and challenging innovations optimizing the talents of faculty and staff and optimizing the benefits of innovative practices and technologies	N/A -2 -1 0 1 2
Act as an advocate and spokesperson for the school to all stakeholders, including parents and community members	N/A -2 -1 0 1 2
Allocate resources in a way that provides faculty with the materials and professional development necessary for successful practice	N/A -2 -1 0 1 2
Acquire knowledge of curriculum, instruction, and assessment practices and use this	N/A -2 -1 0 1 2

19.a. Did participating in a mentoring program help you make the adjustment from your previous position to that of a school leader? ___ Yes ___ No

19.b. Please explain your yes or no answer.

20.a. Did you leave or have you ever considered leaving your position of school leadership?
___ Yes ___ No

20.b. If yes, why?

20.c. If no, did the mentoring program impact your decision not to leave? ___ Yes ___ No

Please explain your yes or no answer.

21. What do you believe has been most useful with helping you to improve teaching and learning in your school? Rank order the top 3 experiences from most helpful (starting with 1).

- | | |
|--|--|
| ___ Principal preparation program | ___ Pre-service Internship |
| ___ Mentoring program | ___ Illinois Administrator Academy courses |
| ___ On-the-job experiences (I learn as I go) | ___ Other professional development experiences |
| | ___ Other (Please explain.) |
-

22. How well prepared for the school leadership position did you feel after completing your principal preparation/licensure program?

- | | | | |
|-----------------|-------------------|---------------------|--------------------|
| Poorly prepared | Somewhat prepared | Adequately prepared | Very well prepared |
| 1 | 2 | 3 | 4 |

23. How well prepared for the school leadership position did you feel after completing the mentoring program?

- | | | | |
|-----------------|-------------------|---------------------|--------------------|
| Poorly prepared | Somewhat prepared | Adequately prepared | Very well prepared |
| 1 | 2 | 3 | 4 |

24. What were the most beneficial components of the mentoring program? (Rank order the top 3 most beneficial components (starting with 1).

- ___ Building a network of support with an advanced colleague
- ___ Building a network of support with peer colleagues
- ___ Learning new leadership practices/strategies from an experienced instructional leader
- ___ Receiving constructive criticism on my ideas and practices
- ___ Receiving emotional support
- ___ Learning the political and social contexts of leading schools
- ___ Receiving administrative and managerial support that helped me prioritize my duties to keep the school's focus on improved student learning
- ___ Other (please explain) _____
- ___ Other (please explain) _____
- ___ Other (please explain) _____

25. What were the most significant barriers confronted in the mentoring program? (Rank order the top 3 most significant barriers (starting with 1).

- Lack of time
- Lack of resources
- Too much distance between me and my mentor
- Communication was difficult to maintain between me and my mentor
- Lack of district support (e.g., superintendent) inhibited my full participation
- Mismatch of professional differences between me and my mentor
- Mismatch on personal differences between me and my mentor
- Mismatch based on gender or race differences between me and my mentor
- Other (please explain) _____
- Other (please explain) _____
- Other (please explain) _____

26. If you could recommend improvements to your mentoring program to make the program more effective for supporting new school leaders, what would they be?

27.a. Finally, would you recommend participating in a mentoring program to a new principal or other school leader?

Yes No

27.b. Please explain your yes or no answer.

Thank you very much for your participation in this survey.

Please place the questionnaire in the enclosed envelope and return to:

**Lisa Hood
Center for the Study of Educational Policy
Campus Box 5900
Normal, IL 61790-5900**

Or, fax the survey to 309/438-8683

Appendix B

Interview Protocol for Clearlake District/Program Administrators

Date & Time: _____ Completed by: _____

Fill out prior to the interview:

Interviewee: _____

Title: _____

Informed Consent Script (to be read prior to the interview)

Hi. My name is [researcher name], and as I/my colleague told you on the phone a couple of weeks ago, I am associated with the Center for the Study of Educational Policy at Illinois State University. My colleagues and I are studying the Clearlake School District's mentoring program for new principals.

During today's interview, I will ask a series of questions. The information I gather from you will be held confidential in which the researchers of this study will not link your data with your name, unless you give me permission to use your name. If I want to use a quote from you, I will ask you to review the quote and will not use it without your permission. You may also decline to answer any questions with which you don't feel comfortable, and you may decide to stop your participation in this interview any time without penalty. Does this sound OK?

_____yes _____no

Would it be OK to tape record the interview so that I don't miss anything?

_____yes _____no

Allow me to begin with a brief preview of what the interview will cover. We will be talking specifically about the evolution of the mentoring program for new principals. Then, I'll ask you to describe the mentoring program as it is being currently implemented. Finally, we'll end our discussion talking about the impacts you think the mentoring program is having on your principals' leadership development, and the subsequent impact this may be having on your districts' schools, teachers, and students.

-
1. Let's go back to the beginning and trace the development of the Clearlake mentoring program. Why was the mentoring program developed?
 - a. Probes:
 - i. What was the impetus?
 - ii. Who was involved in developing the program?
 - iii. What was the process? Where did you go to find a model of what a mentoring program should look like?

 2. What problem or set of conditions is the mentoring program intended to address?

3. How has the mentoring program changed over the years?
4. Why did you or the other program administrators feel like the program needed to be changed?
5. Please describe the mentoring program as it is being currently implemented.
Probes:
 - a. How are mentors selected?
 - b. How are mentors trained?
 - c. How are mentors and protégés matched?
 - d. Are there any requirements for the mentoring process?
 - i. Journals
 - ii. Logs
 - iii. Specific assignments
 - iv. Assessments
6. How does the mentoring fit with the district's overall vision and goals?
7. How does the mentoring program fit into the district's professional development plan for new principals?
8. What factors do you think affect the implementation and effectiveness of the mentoring program? (e.g., mentor characteristics, principal characteristics, school demographic factors, student demographic characteristics, district characteristics)
 - a. Which factors affect the program positively?
 - b. Which factors affect the program negatively?
9. How did you choose this particular model for your district's mentoring program?
10. Why do you think it will have a beneficial impact on principals and your schools?
11. In what ways do you believe that the mentoring program is improving principals' leadership development? (e.g., leadership practices, adjustment to the position, confidence in their ability)
12. Do you think that the changes you are seeing in principals' leadership practices are having an impact on:
 - a. School climate?
 - b. Teachers and their teaching practices?

- c. Student learning?
 - d. Other?
13. Are there further changes you think need to be made to the mentoring program so that it will have greater impact on principals' leadership capacities?
 14. That's all the questions I have for you. Is there anything we did not discuss that you would like to add?
 15. Do you have any questions for me about the study?

Thank you for taking the time for this interview. Your perspective will be very valuable to this study.

Appendix C

Interview Protocol for Clearlake Principal Protégés

Date & Time: _____ **Completed by:** _____

Fill out prior to the interview:

Interviewee: _____
Title: _____
School: _____
School Minority Rate: _____
School Free/Reduced Lunch
Rate: _____

Informed Consent Script (to be read prior to the interview)

Hi. My name is [researcher name], and as I/my colleague told you on the phone/email a couple of weeks ago, I am associated with the Center for the Study of Educational Policy at Illinois State University. My colleagues and I are studying the Clearlake School District’s mentoring program for new principals.

In my message to you asking you to participate in the study, I told you that you were asked to take part in the study because you participated in the Clearlake School District mentoring program. The design of this study is a case study with multiple mini-studies of principals like you who were or are mentoring protégés. For your mini-case study, I will be interviewing you about your experiences with the mentoring program; collecting artifacts that you attribute to your participation in the program; surveying your teachers in regards to school climate, teaching and learning practices and outcomes; and talking with your mentor. I want to reiterate to you that this is NOT an evaluation of you or your abilities as a leader. This is a study of the district’s mentoring program and the impacts it has had on principals and their schools. The information collected during this study will not be used by district administrators in their evaluation of you as a principal. Data reported to district administrators and in other publications will contain pseudonyms in place of your name, and the name of your mentor.

During today’s interview, I will ask a series of questions. The information I gather from you will be held in confidence. When I use it, I will use a pseudonym so that your responses cannot be linked to you. If I want to use a quote from you, I will ask you to review the quote and will not use it without your permission. You may also decline to answer any questions with which you don’t feel comfortable, and you may decide to stop your participation in this interview any time without penalty. Does this sound OK?

_____yes _____no

Would it be OK to tape record the interview so that I don’t miss anything?

_____yes _____no

Allow me to begin with a brief preview of what the interview will cover. I will begin the interview by asking you for background information on your history with the district. Then we will be talking specifically about the mentoring program as you have experienced it. Then, I’ll ask you to describe the mentoring program as it is being currently implemented. Finally, we’ll end our discussion talking about the impacts you think the mentoring program is having on your leadership development, and the subsequent impact this may be having on your school, teachers, and students.

1. First, let's start with some questions about your history in the district. How long have you worked in the Clearlake School District?
2. How long have you been a principal?
3. What other positions have you held in your career as an educator?
4. How long have you been at this school?
5. From which Type 75 preparation program did you earn your degree?
6. In what capacity have you participated in the Clearlake mentoring program?
 - a. Protégé?
 - b. Mentor?
 - c. Other?
7. Describe the mentoring program for me as you have experienced it?
 - a. Describe your mentor (e.g., male/female, years of experience as a principal, years of experience as a mentor, accomplishments as a principal).
 - b. Describe the requirements for the mentoring experiences (e.g., a journal or log, time requirements, specific content addressed).
 - c. Describe the nature of your meetings with your mentor in terms of how you met, how often you met, and what support was offered.
8. Do you believe that the mentor with whom you were matched was an appropriate mentor for you?
 - a. If yes, why was this a good match?
 - b. If no, how could the match have been a better one?
9. Do you think your mentor was adequately trained for the role?
 - a. If yes, why?
 - b. If no, why not?
 - c. How could the training of mentors be improved?
10. (Optional: If the principal also served as a mentor: Do you think that you were adequately trained for the role of mentor?)

- a. If yes, why?
 - b. If no, why not?
 - c. How could the training of mentors be improved?
11. Was there a specific focus that guided your mentoring experiences? In other words, did you and your mentor work to address some specific leadership needs or challenges you were facing in your first years in the principalship?
12. In what way do you believe the mentoring program has improved your leadership development in terms of:
- a. Your practices?
 - b. Your adjustment to the position?
 - c. Your confidence in your ability?
13. (Use the responses to question 11 to probe into how the mentoring program impacted their leadership development in those 3 areas.) Can you expand on how the mentoring program improved your leadership development? What were some specific qualities of the experience that enhanced your leadership?
14. Have you experienced any negative impacts from your participation in the mentoring program? If so, could you describe these?
15. In what areas hasn't the mentoring program addressed your leadership needs?
16. (Optional: if the principal has also served as a mentor). How do you think your role as a mentor has provided continuous professional development?
- a. Are you reaping any benefits from serving as a mentor?
 - b. Are there any pitfalls to serving as a mentor?
17. In what ways do you think the changes you see in your leadership practices that are a result of the mentoring program are having subsequent impacts on:
- a. Your school climate?
 - b. Your teachers?
 - c. Student learning and engagement?
 - d. Other?

18. What contextual factors do you think impact the implementation and effectiveness of the mentoring program? (e.g., mentor characteristics, principal characteristics, school demographic factors, student demographic characteristics, district characteristics).
 - a. Which factors impact the program positively?
 - b. Which factors impact the program negatively?
19. Overall, how do you see the mentoring program fitting in with the professional development plan the district implements with its new principals?
20. Overall, how do you think the mentoring program is impacting the leadership capacity of the Clearlake School District?
21. Are there further changes you think need to be made to the mentoring program so that it will have greater impact on principals' leadership capacities?
22. That's all the questions I have for you. Is there anything we did not discuss that you would like to add?
23. Do you have any questions for me about the study?

Thank you for taking the time for this interview. Your perspective will be very valuable to this study.

Appendix D

Interview Protocol for Clearlake Mentors

Date & Time: _____ **Completed by:** _____

Fill out prior to the interview:

Interviewee: _____

Title: _____

Informed consent script (to be read prior to the interview)

Hi. My name is [researcher name], and as I/my colleague told you on the phone a couple of weeks ago, I am associated with the Center for the Study of Educational Policy at Illinois State University. My colleagues and I are studying the Clearlake School District’s mentoring program for new principals.

During today’s interview, I will ask a series of questions. The information I gather from you will be held in confidence. When I use it, I will use a pseudonym so that your responses cannot be linked to you. If I want to use a quote from you, I will ask you to review the quote and will not use it without your permission. You may also decline to answer any questions with which you don’t feel comfortable, and you may decide to stop your participation in this interview any time without penalty. Does this sound OK?

_____yes _____no

Would it be OK to tape record the interview so that I don’t miss anything?

_____yes _____no

Allow me to begin with a brief preview of what the interview will cover. I will begin the interview by asking you for background information on your history with the district. Then we will be talking specifically about the mentoring program as you have experienced it. Finally, we’ll end our discussion talking about the impacts you think the mentoring program is having on your protégés leadership development, specifically (name of protégé) and the subsequent impact this may be having on his/her school, teachers, and students.

-
1. First, let’s start with some questions about your history in the district. What school leadership positions have you worked in the Clearlake school district? How long have you worked/did you work in the Clearlake School District (depending on current career status)?
 2. How many principals have you mentored?
 3. Describe the mentoring program for me as you have experienced it?
 - a. How would you describe your role as a mentor?
 - b. Describe the requirements for the mentoring experiences (e.g., a journal or log, time requirements, specific content addressed)

- c. Describe the nature of your meetings with your protégés in terms of how you communicated, how often you communicated, and what support was offered.
4. Now, thinking of the protégé (name of principal) with whom you are currently mentoring, do you believe that the protégé with whom you were matched was an appropriate match for the protégé?
 - a. If yes, why was this a good match?
 - b. If no, how could the match have been a better one?
5. Do you think you were adequately trained for the mentoring role?
 - a. If yes, why?
 - b. If no, why not?
 - c. How could the training of mentors be improved?
6. Was there a specific focus that guided your mentoring experiences? In other words, did you and your protégé work to address some specific leadership needs or challenges he/she was facing in the first years in the principalship?
7. (If the mentor is a practicing administrator): How do you think your role as a mentor has provided continuous professional development for your own practice?
 - a. Are you reaping any benefits from serving as a mentor?
 - b. Are there any pitfalls to serving as a mentor?
8. How do you think the mentoring program is improving your protégés' leadership practices in terms of:
 - a. His/her leadership practices?
 - b. His/her adjustment to the position?
 - c. His/her confidence in their leadership abilities?
9. Are you seeing that the mentoring program is having subsequent impacts on the protégés' school in terms of :
 - a. School climate?
 - b. Teachers?
 - c. Student learning and engagement?
 - d. Other?

10. What contextual factors do you think impact the implementation and effectiveness of the mentoring program? (e.g., mentor characteristics, principal characteristics, school demographic factors, student demographic characteristics, district characteristics).
 - a. Which factors impact the program positively?
 - b. Which factors impact the program negatively?
11. Overall, how do you see the mentoring program fitting in with the professional development plan the district implements with its new principals?
12. Overall, how do you think the mentoring program is impacting the leadership capacity of the Clearlake School District?
13. Are there further changes you think need to be made to the mentoring program so that it will have greater impact on principals' leadership capacities?
14. That's all the questions I have for you. Is there anything we did not discuss that you would like to add?
15. Do you have any questions for me about the study?

Thank you for taking the time for this interview. Your perspective will be very valuable to this study.

Appendix E

Self-Reflection Protocol

Directions: Review the reflection journal that contains my journal entries at different points in the evaluation design, implementation, analysis and reporting. Code the journal according to the quality and utility criteria as well as the benefits and challenges I experienced during the evaluation.

Evaluation Quality

1. Credibility:

- a. a logical link between the data and the program theory developed to represent the mentoring program
- b. the account of the program seems plausible and empirically grounded in the data available?

2c. How could I have done a better job of improving the linkage between the data and the program theory? Are there any data that appears to be missing?

What were the benefits of integrating constructivism and program theory evaluation that affected providing credible results?

What were the challenges or tensions from integrating the two theories that affected providing credible results?

3. Context-Program Interactions: The program theory adequately presented the relationship between the program and its context?

What were the benefits of integrating constructivism and program theory evaluation that affected presenting the context-program interactions?

What were the challenges or tensions from integrating the two theories that affected presenting the context-program interactions?

4. Fairness: Include the multiple perspectives of program stakeholders? (*fairness*)

4b. Did I privilege some perspectives over others? Whose perspectives seemed most privileged? Does this seem appropriate? Whose perspectives should have been included more in the theory?

What were the benefits of integrating constructivism and program theory evaluation that affected the fairness of the evaluation?

What were the challenges or tensions from integrating the two theories that affected the fairness of the evaluation?

When bringing in the multiple perspectives, what were the benefits and challenges/tensions that affected the C-PTE?

5. Principled Discovery: To what extent did is there evidence that I/the evaluator systematically investigated program processes, context, and effects to develop the program theory using deductive and inductive reasoning

5b. Competitive Elaboration: Identify alternative explanations to the evaluation's findings and represent their relationship to the final program theory

How did integrating the two theories help uncover alternative explanations? (*competitive elaboration*)

How did using inductive and deductive data collection and analysis strategies influence program theory development? Conversely, how did program theory development influence the data collection and analysis strategies? (*principled discovery*)

Evaluation Utility

6. Resonance: Capturing the program in an overall sense, providing rich description (i.e., do you feel like you understand what the program is, the actors involved, the intended purpose and goals, and so forth)?

What were the benefits of integrating constructivism and program theory evaluation that affected the resonance of the evaluation results?

What were the challenges or tensions from integrating the two theories that affected providing the resonance of the evaluation results?

7. Resonance and Transferability: To what extent does the program theory as a representation (or portrait) of this new principal mentoring program resonate with your experiences with similar programs?

What were the benefits of integrating constructivism and program theory evaluation that affected the resonance and transferability of the evaluation?

What were the challenges or tensions from integrating the two theories that affected the resonance and transferability of the evaluation?

9. Information Scope and Selection: Data collected in this evaluation broad enough to explore pertinent questions about the program that is useful to the field?

What were the benefits of integrating constructivism and program theory evaluation that affected the scope of the evaluation?

What were the challenges or tensions from integrating the two theories that affected the scope of the evaluation?

10. Values Identification: The perspectives, rationale, and procedures used in order to make judgments of program value

a. Evidence of bias?

What were the benefits of integrating constructivism and program theory evaluation that affected making appropriate judgments of program value?

What were the challenges or tensions from integrating the two theories that affected making appropriate judgments of program value?

Appendix F

Program Evaluators Interview Guide

Informed consent information:

Your participation in this interview is voluntary and you can stop the interview at any time without worry of penalty. Your responses will be held confidential and I will not identify you in my dissertation or any subsequent publications.

I would like to record our discussion today to ensure the accuracy of my notes. Is that okay?

_____yes _____no

If at any time you would like me to stop the recording, let me know and I will turn off the recorder.

- 1. Before we begin, are there any questions about the evaluation—the evaluation process, the report, the school district, or the mentoring program that you need addressed before we begin the questions?**
- 2. To begin this discussion, I defined this evaluation as a constructivist-program theory**

Now I would like to begin my metaevaluative questions starting with the evaluation quality. If at any point you need more information about the evaluation in order to help you answer the questions, please don't hesitate to ask.

Evaluation Quality

- 3. To what extent is there a logical link between the data that was collected and presented in the report and the program theory developed for the mentoring program? (*credibility*)**

To no extent To some extent To great extent

- 3b. To what extent does the account of the program seem believable and supported by the data presented in the report?**

To no extent To some extent To great extent

- 3c. How could the evaluator, have improved the linkage between the data and the program theory? Do any data appear to be missing? Did any of the data seem unclear or irrelevant to explain how the mentoring program works or how effective it is?**

- 4. To what extent did the program theory adequately explain the relationship between the program and the Clearlake district in terms of its beliefs, goals, and practices? (*context-program interactions*)**

To no extent To some extent To great extent

- 4b. How could the evaluator have better described the relationship between the program and the district? How could the program theory be modified to represent this relationship? How important is it to relate the program to its context in order to better understand the way a program works and its effectiveness?**

5. To what extent did the program theory appear to include the multiple perspectives of program stakeholders? (*fairness*)

To no extent

To some extent

To great extent

5b. If you answered “to no or to some extent”, whose perspectives seemed most prominent? Does this seem appropriate? Whose perspectives should have been included more prominently in the evaluation findings?

6. To what extent did you see evidence that the evaluator systematically investigated the mentoring program’s processes, context, and effects to develop the program theory? (*principled discovery*)

To no extent

To some extent

To great extent

6b. Sometimes it is possible that there are other alternative explanations that could account for the evaluation findings. To what extent did the evaluator identify possible alternative explanations and describe their effects on the final program theory? (*competitive elaboration*)

To no extent

To some extent

To great extent

6c. If you answered “to no or to some extent”, in what ways could the evaluator have ruled out the alternative findings or described the relationship to the final program theory that would have been more believable?

Evaluation Utility

7. Overall, to what extent did the evaluator capture the essence of the program (i.e., do you feel like you understand what the program is, the actors involved, the intended purpose and goals, and so forth)?

To no extent

To some extent

To great extent

7b. If you answered “to no or to some extent”, what information should have been included to develop a richer description?

8. To what extent did the evaluator describe the criteria and data she used to make judgments about the mentoring program’s value? (*values identification*)

To no extent

To some extent

To great extent

8b. What information should have been provided to make the evaluator’s judgments clearer? Did you notice a bias? If so, explain.

The following questions are to rule out any possible biases in your responses to the quality and usefulness of the program theory evaluation of the new principal mentoring program due to an unclearly written report.

9. To what extent was the report clear and easy to understand in terms of its descriptions of the program, the evaluation process, and the evaluation findings? (*report clarity*)

To no extent

To some extent

To great extent

9b. If you answered “to no or to some extent”, what was unclear? Did this influence your answers to any of the questions asked above?

10. Are there any other questions that I should have asked that would add to this discussion on evaluation quality and usefulness? In other words, are there any other quality or use criteria that I should have added?

11. For this study, I would like to compare my constructivist program theory evaluation with exemplars of program theory evaluation, and/or an evaluation that you believe is a good example of a constructivist program theory evaluation. I will compare structures, content, and compare the reports on the quality and utility criteria that I used in our interview. Can you suggest one or two evaluations that I might use as a comparison? (Ask them where I can find it, if the report is publicly available or if they can furnish a version for me.)

Thank you for engaging in this discussion with me today. Your reflections will make a valuable contribution to my study.

Appendix G

Mentoring Program Administrators Interview Guide

Informed consent information:

Your participation in this interview is voluntary and you can stop the interview at any time without worry of penalty. Your responses will be held confidential and I will not identify you in my dissertation or any subsequent publications.

I would like to record our discussion today to ensure the accuracy of my notes. Is that okay?

_____yes _____no

If at any time you would like me to stop the recording, let me know and I will turn off the recorder.

2. Before we begin, are there any questions about the evaluation—the evaluation process, the report, the school district, or the mentoring program that you need addressed before we begin the questions?

Now I would like to begin my questions on the evaluation starting with questions related to evaluation quality. If at any point you need more information about the evaluation in order to help you answer the questions, please don't hesitate to ask.

Evaluation Quality

3. To what extent is there a logical link between the data that was collected and presented in the report and the program theory developed for the mentoring program?

To no extent To some extent To great extent

2b. To what extent does the account of the program seem believable and supported by the data presented in the report?

To no extent To some extent To great extent

2c. How could the evaluator have improved the linkage between the data and the program theory? Do any data appear to be missing? Did any of the data seem unclear or irrelevant to explain how the mentoring program works or how effective it is?

3. To what extent did the program theory adequately explain the relationship between the program and the Clearlake district in terms of its beliefs, goals, and practices?

To no extent To some extent To great extent

3b. How could the evaluator have better described the relationship between the program and the district? How could the program theory be modified to represent this relationship? How important is it to relate the program to its context in order to better understand the way a program works and its effectiveness?

4. To what extent did the program theory appear to include the multiple perspectives of program stakeholders?

To no extent To some extent To great extent

4b. If you answered “to no or to some extent”, whose perspectives seemed most prominent? Does this seem appropriate or did you notice a bias? Whose perspectives should have been included more prominently in the evaluation findings?

5. To what extent did you see evidence that the evaluator systematically investigated the mentoring program’s processes, context, and effects to develop the program theory?

To no extent To some extent To great extent

5b. Sometimes it is possible that there are other alternative explanations that could account for the evaluation findings. To what extent did the evaluator identify possible alternative explanations and describe their effects on the final program theory?

To no extent To some extent To great extent

5c. If you answered “to no or to some extent”, in what ways could the evaluator have ruled out the alternative findings or described the relationship to the final program theory that would have been more believable?

Evaluation Utility

6. Overall, to what extent did the evaluator capture the essence of the program (i.e., do you feel like you understand what the program is, the actors involved, the intended purpose and goals, and so forth)?

To no extent To some extent To great extent

6b. If you answered “to no or to some extent”, what information should have been included to develop a richer description?

7. To what extent does the program theory as a representation (or portrait) of this new principal mentoring program reflect your own experiences with new principal mentoring programs and your beliefs about how and why they work?

To no extent To some extent To great extent

7b. If “to great extent”, how is this account of the Clearlake mentoring program similar to your experiences and beliefs?

7c. If “to no or to some extent”, how does this account of the Clearlake mentoring program differ from your experiences and beliefs?

8. To what extent does this program theory of the Clearlake mentoring program offer you deeper insights into your work as a mentoring program administrator?

To no extent

To some extent

To great extent

8b. If “to great extent”, what did you learn from this account of the Clearlake mentoring program that will help you administer your current or future principal mentoring programs?

8c. If “to no or to some extent”, what information would you have liked to have seen in this account of the Clearlake mentoring program that would have helped you administer your current or future principal mentoring programs?

8d. What information do you typically use or find helpful when you are examining the effectiveness of your own mentoring program and wanting to make improvements? What information is most helpful?

8e. What in the report was the most helpful? Was it the graphical representation of the program? The narrative description? The combination of the both? Would the report have been more or less helpful without either of these 2 components?

9. To what extent were the data collected in this evaluation broad enough to explore important and relevant issues about mentoring programs that is useful to the field?

To no extent

To some extent

To great extent

9b. What other evaluation questions should have been asked? What other data would have added to the evaluation’s usefulness?

10. To what extent did the evaluator describe the criteria and data she used to make judgments about the mentoring program’s value?

To no extent

To some extent

To great extent

10b. What information should have been provided to make the evaluator’s judgments clearer?

The following questions are to rule out any possible biases in your responses to the quality and usefulness of the program theory evaluation of the new principal mentoring program due to an unclearly written report.

11. To what extent was the report clear and easy to understand in terms of its descriptions of the program, the evaluation process, and the evaluation findings?

To no extent

To some extent

To great extent

11b. If you answered “to no or to some extent”, what was unclear? Did this influence your answers to any of the questions asked above?

12. Are there any other questions that I should have asked that would add to this discussion on this evaluation’s quality and usefulness?

In fairness—should I have disclosed my familiarity with the program—speaks to issue of bias

13. Finally, for this study, I would like to compare my evaluation with other research or evaluation reports that ones that you have found particularly useful in informing your work of administering your principal mentoring program (or other leadership professional development programs). Can you suggest one or two reports that I might use as a comparison? (Ask them where I can find it, if the report is publicly available or if they can furnish a version for me.)

Thank you for engaging in this discussion with me today. Your reflections will make a valuable contribution to my study.

Appendix H

Description of Evaluations Used for Comparative Analysis

Bickman, L. (1996). The application of program theory to the evaluation of a managed mental health care system. *Evaluation and Program Planning*, 19(2), 111-119.

Bickman and his colleagues conducted a program theory evaluation of the Ft. Bragg Child and Adolescent Mental Health Demonstration managed care system of mental health services for children in military families. In this new system, children and adolescents received a continuum of services that administrators believed would improve treatment outcomes as well as control costs. Bickman conducted a program theory evaluation to test the administrators' theory about how the program should work and whether it achieved its outcomes. The evaluators developed an a priori program theory of how the program should work, and then tested the theory through interviews with program administrators and staff and document reviews. The design of this evaluation was a quasi-experimental design in which the comparison group was children and adolescents on military bases without the new managed care system whose treatment and outcomes were compared to children and adolescents on military bases that were implementing the new system. The outcomes for clients were measured through an analysis of the clients' medical records in which evaluators were looking for improvements in the efficiency in the intake and treatment process, parent satisfaction with the services, improvement in mental health symptoms, lower costs per client, higher client satisfaction with the services, and faster recovery from symptoms.

Bowen, K.A. (1999). *Evaluation of a teen pregnancy and parenting program: Integrating an interpretivist, theory-oriented approach* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database. (AAT 9923763).

Bowen conducted an interpretivist program theory evaluation of the Teen Pregnancy and Parenting Program. The purpose of this program was to prevent teenage pregnancy and provide health, education, nutrition and other social services to parenting teenagers. The focus of this evaluation was to develop a program theory that describes the program's context, processes, and outcomes. The evaluator developed this theory through interviews with program staff and clients. Bowen also conducted a concept mapping exercise with staff members as well as with the teenage clients to elicit their views on the program's most important activities followed up with staff and client interviews. The evaluator also observed the program staff working with clients in their homes. The evaluation ended with a normative theory of the program as implemented.

Crane, B. (2000). *Building a theory of change and a logic model for an empowerment-based family support training and credentialing program* (Doctoral dissertation). Retrieved from ProQuest Dissertation and Theses Database. (AAT9967470).

This evaluation was an interpretivist program theory evaluation of the New York State Family Development Training and Credentialing Program. This program provides training to staff members at community action agencies in ten core competencies (e.g., building relationships with families, family self-reliance, home visiting). Crane developed a program theory through interviews, focus groups, and meetings with training staff, community action staff members who had participated in the training, and family members who had sought services at the community organizations. Through these data collection sources, Crane was able to develop a program theory that identified the program's essential inputs, primary activities, and finally the short-, intermediate, and long-term outcomes. For each of these elements of the program theory, Crane substantiated her interpretations that given inputs of funding, curriculum, and state agency support, trainers are able to provide the training to community action personnel in the ten competencies who then put these competencies in action in their work with families. As a result, families and their children have better life outcomes in stable incomes, youth engagement in education and community activities, and family empowerment.

Ford, R. (1995). *A non-traditional evaluation of the Woodbury Public School Elementary Enrichment Program* (Doctoral dissertation). Retrieved from ProQuest Dissertation and Theses Database. (AAT 9527469).

Ford conducted a constructivist evaluation of the Woodbury Elementary Gifted Program that was being evaluated at the request of the school district. The school board and district administrators were trying to make a decision about the future of the program—whether it would remain in the school as implemented or if it would be modified. To conduct this evaluation, Ford conducted interviews with parents with students in the gifted program and parents whose children were not in the program, school board members, teachers in the program, teachers in general education classrooms, and school and district administrators. The evaluator attended and observed many meetings where the fate of this program was debated. The evaluation report was clear in describing the political tensions inherent in this difficult decision that the district was deliberating. In the report, it became obvious that the evaluators became not just observers of the process but participants in a committee that developed proposals to submit to the school board which recommended designing and implementing enrichment activities throughout the K-5 grades rather than continue the self-contained gifted classroom.

Marchal, B., Dedzo, B. & Kegels, G. (2010). A realist evaluation of the management of a well-performing regional hospital in Ghana. *BMC Health Services Research*, 10(24), <http://www.biomedcentral.com/1472-6963/10/24>. Retrieved September 28, 2010.

This was a realist evaluation of the human resource management practices in a hospital that was considered well-performing in Ghana. The evaluators were interested in using a realist evaluation model to study these practices with the belief that this evaluation approach would more accurately identify the features of the management practices (e.g., induction programs for staff, training, and improved communication practices) that contributed to the successful

performance of the hospital. Evaluators conducted interviews, focus groups, observations, and document reviews to develop context-mechanism-outcome configurations that would explain which services were having what outcomes and under what conditions.

Appendix I

Comparative Analysis Protocol

4. Do I agree with how they defined the program evaluation approach the evaluator used? How is this evaluation different from and similar to my evaluation?

Evaluation Quality

5. To what extent is there a logical link between the data that was collected and presented in the report and the program theory developed for the mentoring program?
(*credibility*)

To no extent To some extent To great extent

3. To what extent does the account of the program seem believable and supported by the data presented in the report?

To no extent To some extent To great extent

- 3b. How could the evaluator, have improved the linkage between the data and the program theory? Do any data appear to be missing? Did any of the data seem unclear or irrelevant to explain how the mentoring program works or how effective it is?

-
4. To what extent did the program theory or narrative in the evaluation report adequately explain the relationship between the program and its context? (e.g., beliefs, goals, and practices) (*context-program interactions*)

To no extent To some extent To great extent

- 4b. How could the evaluator have better described the relationship between the program and its context? How could the program theory or evaluation findings be modified to represent this relationship? How important is it to relate the program to its context in order to better understand the way a program works and its effectiveness?

-
5. To what extent did the program theory or evaluation findings appear to include the multiple perspectives of program stakeholders? (*fairness*)

To no extent To some extent To great extent

- 5b. If I answered “to no or to some extent”, whose perspectives seemed most prominent? Does this seem appropriate? Whose perspectives should have been included more prominently in the evaluation findings?

6. To what extent did I see evidence that the evaluator systematically investigated the program's processes, context, and effects to develop the program theory/evaluation findings? (*principled discovery*)

To no extent

To some extent

To great extent

6b. Sometimes it is possible that there are other alternative explanations that could account for the evaluation findings. To what extent did the evaluator identify possible alternative explanations and describe their effects on the final program theory/evaluation findings? (*competitive elaboration*)

To no extent

To some extent

To great extent

6c. If I answered "to no or to some extent", in what ways could the evaluator have ruled out the alternative findings or described the relationship to the final program theory/findings that would have been more believable?

Evaluation Utility

7. Overall, to what extent did the evaluator capture the essence of the program (i.e., do you feel like you understand what the program is, the actors involved, the intended purpose and goals, and so forth)?

To no extent

To some extent

To great extent

7b. If I answered "to no or to some extent", what information should have been included to develop a richer description?

Process captured well

8. To what extent did the evaluator describe the criteria and data s/he used to make judgments about the program's value? (*values identification*)

To no extent

To some extent

To great extent

8b. What information should have been provided to make the evaluator's judgments clearer? Did I notice a bias? If so, explain.

8d. What in the report was the most helpful? (e.g., Was it a graphical representation of the program? The narrative description? The combination of the both? Other?)

10. To what extent was the report clear and easy to understand in terms of its descriptions of the program, the evaluation process, and the evaluation findings? (*report clarity*)

To no extent

To some extent

To great extent

10b. If I answered “to no or to some extent”, what was unclear? Did this influence my answers to any of the questions asked above?

Appendix J

Criteria Used to Judge the Quality of the District Mentoring Program

The Wallace Foundation (2007) issued a report on new principal mentoring programs that listed criteria for quality mentoring programs. They included:

1. High quality training for mentors
2. Data collection on program efficacy (i.e., how the program is (or is not) building leadership behaviors and dispositions)
3. Provide for at least one year of mentoring, two years is ideal
4. Funding that is sufficient to provide mentor stipends, and
5. Clear program goals that are focused on providing new principals with the knowledge and skills to provide leadership directed at improving teaching and learning

These criteria were used to evaluate the quality of the program, along with the administrators', protégés', and mentors' perspectives on the worth and satisfaction with the mentoring program in supporting new school leaders in the early years to become effective leaders of their schools.