A CASE STUDY EVALUATING THE FIDELITY OF IMPLEMENTATION OF CONSTRUCTING MEANING TRAINING AT A LOCAL MIDDLE SCHOOL

by

BRIAN F. SICA

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

March 2016

DISSERTATION APPROVAL PAGE

Student: Brian F. Sica

Title: A Case Study Evaluating the Fidelity of Implementation of Constructing Meaning Training at a Local Middle School

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Methodology, Policy, and Leadership by:

Keith Zvoch	Chairperson
Joanna Smith	Core Member
Yvonne Curtis	Core Member
Audrey Lucero	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded March 2016

DISSERTATION ABSTRACT

Brian F. Sica

Doctor of Education

Department of Educational Methodology, Policy, and Leadership

March 2016

Title: A Case Study Evaluating the Fidelity of Implementation of Constructing Meaning Training at a Local Middle School

The purpose of this study was to understand the implementation of practices derived from Constructing Meaning (CM) training by teachers (*n* = 30) at a local middle school. The study took place in two phases. Phase one was primarily quantitative. Implementation fidelity was measured for each critical component of CM training, and component and aggregate indices were constructed and analyzed. The second phase, primarily qualitative, investigated teachers' perceptions of the conditions that favored or hindered implementation. Results indicated that certain components were implemented to a greater degree than others and that the overall implementation fidelity was approximately 50%. Key conditions for implementation were identified as collaboration (both with peers and CM trainers), sufficient time, and clear connections to other programs.

CURRICULUM VITAE

NAME OF AUTHOR: Brian F. Sica

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED

University of Oregon, Eugene

Montana State University, Bozeman

University of Idaho, Moscow

DEGREES AWARDED

Doctor of Education, Educational Methodology, Policy, and Leadership, 2016, University of Oregon

Master of Science, Science Education, 2006, Montana State University

Bachelor of Science, Chemistry Teaching, 2001, University of Idaho

AREAS OF SPECIAL INTEREST

Minority and underrepresented student education

Implementation of educational initiatives, particularly at the classroom level

Adult leadership

PROFESSIONAL EXPERIENCE

Principal, Beaverton School District’s Health and Science School and School of Science and Technology, 2014–present

Assistant Principal, Hillsboro School District’s Century and Hillsboro High Schools, 2010–2014

Chemistry and Physics Instructor, Hillsboro School District’s Hillsboro High School, 2006–2010

Chemistry and Physics Instructor, Shelley School District’s Shelley High School, 2001, 2006

GRANTS, AWARDS, AND HONORS

CTE Revitalization Grant ($362,000), Project Director, Health and Science School, 2014–2015

United States Department of Education Smaller Learning Communities Grant (1.25M), Project Manager, Hillsboro High School, 2008–2011

PUBLICATIONS

Sica, B. (2009, October 31). *Modifying discussion and assessment techniques to increase student understanding and teacher reflective practices.* Retrieved from Action Research Expeditions website: www.arexpeditions.montana.edu. Available March 2010.

ACKNOWLEDGMENTS

I would like to offer my sincere appreciation to Drs. Zvoch, Smith, Lucero, and Curtis for their guidance throughout this project. In addition, a special thanks to the administration, teachers, and students at the middle school site of this project. Also, I would like to thank my cohort members at the University of Oregon; I learned more from you than I ever thought possible. Finally, I would like to thank my wife and son for their unwavering support and sacrifice.

This project is dedicated to my students, past and present. I think of you often.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Federal and local pressure to produce measurable increases in student achievement remains a constant focus for schools across the country (Polikoff, McEachin, Wrabel, & Duque, 2013). In the Race to the Top competitive grant program, states and local educational agencies (LEAs) competed for $4.35 billion in federal grants to be used to improve their schools (Race to the Top Act of 2011, 2014). Virtually every aspect of the detailed application criteria was in some way tied to measureable student achievement. Similarly, beginning in 2011, states were able to apply for flexibility waivers to the Elementary and Secondary Education Act, specifically in regard to the student achievement requirements of the 2001 reauthorization known as No Child Left Behind (NCLB; No Child Left Behind, 2001). As of January of 2016, 43 states have approved requests for waivers. Each of these requests were required to include a detailed plan for improving instruction and closing the achievement gap, as measured by standardized test scores (U.S. Department of Education, 2012).

In efforts designed to meet accountability requirements, many districts have focused on improving curriculum, instruction, and assessment through high quality professional development (Blank & de las Alas, 2009; Darling-Hammond & Wei, 2009). Professional development is usually targeted at an initiative or intervention aimed at a particular curricular area such as literacy or math, or a specific group of students, such as English Language Learners (ELLs). Common professional development initiatives include Professional Learning Communities (PLCs), Positive Behavior Interventions and Supports (PBIS), and programs aimed at increasing the academic English language development of students (Echevarria, Richards-Tutor, Chinn, & Ratleff, 2011). Professional development (PD) can be delivered in various formats. Generally, the formats can be classified as workshop-style, visits to other sites, coaching, research, and peer-to-peer observations (Darling-Hammond & Wei, 2009). Typical professional development includes some combination of the formats, such as initial training, release time for teachers to create and modify curricula, and instruction on the use of program materials (Odden, Archibald, Fermanich, & Gallagher, 2012). It has been estimated that approximately 90% of teachers experience some sort of PD in a given school year, and that 90% of the PD teachers participate in is primarily organized on a workshop model where they attend a one- to three-day conference with little to no systematic follow-up (Darling-Hammond & Wei, 2009). However, the workshop model has not been shown to be the most effective form of PD (Gulumhussein, 2013). In a comprehensive meta-analysis of more than 1,300 studies, Garet et al. (2009) determined the highest effect sizes were from PD formats that were “sustained and intensive” (p. 938). The researchers went on to suggest that models with less than 14 hours of direct instruction had no effect on student achievement (Garet, Porter, Desimone, Birman, & Yoon, 2009).

The difficulty in designing PD to effect change may come less from the specific method used to teach the teachers and more from the level of implementation planning provided (Gulumhussein, 2013). It has been suggested that the challenges of changing practice do not come with practitioners learning the new practice, but rather in their attempts to integrate it into their regular routines (Guskey, 2002). It may take a teacher more than 20 attempts at implementing a practice to master it (Joyce & Showers, 2002). The challenge of implementation is compounded by the desire of school leaders, who likely feel pressure to maximize their resource allocations and see immediate results (Gulumhussein, 2013).

Calculating the exact cost of PD is difficult due to the variety of resources used for implementation. For example, in addition to the cost of training, many initiatives require the development of materials and additional planning time for the teaching staff. Some researchers have estimated that a school district spends between 2% and 5% of its operating budget on PD (Miller, Lord, & Dorney, 1994; Odden et al., 2012). This estimate of the financial impact may be low, however, because it is difficult to assess accurately the amount of time—both compensated and uncompensated—that teachers allocate to implementation of PD skills (Odden et al., 2012). These costs are coming at a time when school and district leaders are forced to balance developmental costs with shrinking budgets. In the 2013–2014 school year, approximately 35 states had lower per-pupil spending than pre-recession levels (Leachman & Mai, 2014).

The combination of budgetary constraints and political pressures to increase achievement outcomes means that school leaders are required to constantly evaluate their programs in order to show a rapid return on investment (ROI). The program evaluations can serve as evidence of ROI if they reveal an improvement in instruction, an increase in student achievement, or both. In order to make inferential claims of improvement, school leaders must design program evaluations using experimental or tightly controlled quasi-experimental designs that include both control and program-receiving (experimental) groups (Weiss, Bloom, & Brock, 2013). However, evaluations are often completed by measuring change in school-wide or district levels of student achievement, typically from standardized tests without the benefit of a solid research design (Shymansky, Wang, Annetta, Yore, & Everett, 2010). Even if an intervention has been previously shown to be effective, evaluating a program by only looking at student achievement data is flawed because it assumes the program has been implemented in a way that would lead to certain expected changes. As a result, in addition to a strong inferential research design, evaluations should also include a measurement of implementation fidelity (Century, Cassata, Rudnick, & Freeman, 2012; Weiss et al., 2013; Zvoch, 2012).

The concept of implementation (or treatment) fidelity, which considers the degree to which a program is delivered as intended (Yeaton & Sechrest, 1981), served as the basis for this case study. The inclusion of treatment fidelity strengthens a program evaluation by giving providers formative data as well as more accurate summative claims (Gulumhussein, 2013; Weiss et al., 2013). In formative evaluations, providers can allocate additional resources or make adjustments to their implementation plans. In summative evaluations, evidence of high implementation fidelity can strengthen inferential claims by demonstrating that the treatment group received the intervention as intended, and was thus distinct from the control group—in other words, the change measured was the result of the intervention (Weiss et al., 2013). Evaluators making inferential claims without a measurement of fidelity risk attributing a change in outcomes to a change in practice that was not verified to have actually occurred (Dusenbury, Brannigan, Falco, & Hansen, 2003).

The study presented here investigated the *manner in which* school leaders and teachers evaluated and understood the degree to which a program had been implemented as intended. The following chapters (1) review and synthesize the relevant literature regarding the concept of fidelity of implementation, (2) describe methods for evaluating implementation using both quantitative and qualitative methods, (3) present findings from an evaluation of Constructing Meaning practices at a local middle school, (4) offer conclusions drawn from the data, and (5) discuss recommendations for application and future research.

CHAPTER II

LITERATURE REVIEW

The literature relevant to this study is reviewed and synthesized in this chapter. The primary themes of the literature review are as follows: a conceptual framework of the construct of fidelity of implementation, a synthesis of the approaches in measuring implementation fidelity in prior research, and the review of the specific PD model being used as an intervention.

Defining Fidelity of Implementation

The concept of fidelity of implementation can be defined as the degree to which a treatment is delivered as intended by its developers (Moncher & Prinz, 1991; Orwin, 2000; Yeaton & Sechrest, 1981). In a research context, evaluating fidelity can provide confirmation that the manipulation of the independent variable occurred as planned (Moncher & Prinz, 1991). In a program monitoring context, fidelity evaluation can provide information to policymakers that services are being implemented as prescribed to reach the intended targets (Orwin, 2000). Although these descriptions seem simple, in practice fidelity of implementation is challenging to define and measure (Zvoch, 2009, 2012).

In their frequently cited study, Dane and Schneider (1998) suggested that fidelity investigations should address five aspects of implementation. *Adherence* is the extent to which the intervention is delivered by a provider as designed by its developer, possibly measured by observations and/or checklists (Drake et al., 2001). In educational settings, the provider is likely a teacher, counselor, or other specialist. *Exposure* (also referred to as *dose*) is a multifaceted construct. Dose is, generally, the completeness of the delivery of the program (Dusenbury et al., 2003). The completeness of delivery can mean the amount of intervention actually received by intended recipients, and is influenced by the methods of delivery and the engagement of recipients. *Quality of delivery* looks at aspects of the intervention beyond basic implementation. Quality of delivery evaluation points can include provider (teacher) enthusiasm, depth of providers' understanding of the program model, and appropriateness of specific applications. *Participant responsiveness* is the level to which the participants (in the case of educational interventions, students) respond to or interact with the intervention. For example, investigators can observe whether a student actually uses the vocabulary list a teacher has posted on the wall. *Program differentiation* documents the degree to which the treatment intervention differs from current practice or a control condition.

Dane and Schneider (1998) suggest that all fidelity studies should measure each of these aspects, though few studies have been able to thoroughly address all five in their evaluations (Dusenbury et al., 2003). Challenges are present in obtaining and utilizing reliable and valid measures of adherence and quality (Dusenbury et al., 2003). For example, dose requires recording every instance of program use, which is only practical through the self-reporting of providers and recipients and introduces the potential for bias and over-reporting (Kruger & Dunning, 1999). Moreover, participant responsiveness can measure a range of recipient actions. Broadly, it may also be measured as simply the number of recipients being presented with the intervention. In a school setting, the number of students in a class that is observed following protocol may all count toward participant responsiveness. A more complete measurement, however, would be a calculation of the number of students actually engaging with the tools of the intervention.

In addition, measuring engagement on a continuum can be very challenging, as it requires observers to interpret varying levels of engagement in different students who are displaying similar actions (Tan, Sun, & Khoo, 2014). For example, a student who appears to be writing may be authentically engaged in a prescribed exercise (high participant responsiveness), while another student who is also writing may be simply writing a message to a friend (low participant responsiveness). Dose can be estimated through the self-reporting of providers and recipients, although the level of bias and over-reporting may be difficult to assess. Measuring program differentiation can be challenging, in that it is common to find similar elements in varying interventions (Hansen, Graham, Wolkenstein, & Rohrbach, 1991). Although the aspects described above can be challenging to accurately measure, they cannot be ignored. Each one represents an important component of the analysis of fidelity of implementation of an intervention.

The terms described by Dane and Schneider are found throughout the literature to introduce and describe fidelity measurement (Carroll et al., 2007; Century, Rudnick, & Freeman, 2010; Dusenbury et al., 2003; Zvoch, 2012). However, they have not been accepted as the standard by all (Weiss et al., 2013). Although frequently referenced in the literature, Dane and Schneider’s terminology has been shown to be too broad to use as a framework of study for implementation fidelity. As described above, the difficulties in measurement have prevented the terminology from being used as a universal framework for fidelity studies.

Why Study Fidelity of Implementation?

The use of interventions to improve outcomes is not unique to the field of education; virtually all service providers implement interventions to change outcomes

(Durlak & DuPre, 2008; Dusenbury et al., 2003). However, early research of implementation fidelity suggested that without studying fidelity of implementation, intervention research does not yield meaningful claims (Yeaton & Sechrest, 1981). In other words, if implementation fidelity is not clearly measured, it is impossible to distinguish between a flawed program and poor implementation. Evaluators must identify whether the intended aspects of the intervention are being fully implemented and delivered to their recipients. Too often, interventions are evaluated based only on the intended outcomes, with little to no measurement of the actual implementation (Dobson & Cook, 1980; Durlak & DuPre, 2008; Harn, Parisi, & Stoolmiller, 2013). Without proper attention to fidelity of implementation, claims made from such evaluations may not accurately reflect the intervention’s actual efficacy.

Fidelity as a summative evaluation. Generally, most practitioners assume that demonstrating high fidelity to evidence-based best practices will result in higher gains in student achievement than those with low fidelity (Harn et al., 2013). However, causal claims regarding effects of an intervention should not be made without including a confirmation of the level of implementation fidelity to complement a well-designed experimental study (Dusenbury et al., 2003; Weiss et al., 2013; Yeaton & Sechrest, 1981). Weiss et al. (2013) proposed a framework for program evaluations that includes the investigation of implementation fidelity with a strong research design. As illustrated in figure 1, Weiss and colleagues describe phase one of their framework as an investigation of fidelity within an experimental design, in order to limit possible errors in interpreting their final outcomes. For example, if fidelity is not measured and student achievement goals are not observed, evaluators may conclude prematurely that the intervention itself was not effective in producing the desired outcomes. Alternatively, when fidelity is measured, researchers can strengthen arguments that the treatment had a causal relationship with reaching desired outcomes by ensuring that the treatment group received the intervention as intended (Echevarria et al., 2011; Wolery, 2011).

Weiss and colleagues describe a comprehensive approach to program evaluation that goes beyond the implementation phase. Their framework includes investigations of the characteristics of the providing organization, characteristics of the recipients, and description of an appropriate experimental or quasi-experimental design. The experimental design phase includes a measurement of treatment contrasts that define and describe the differences between treatments received with and without access to the intervention. The model also includes “mediators” as an intermediary between the treatment being received and the outcomes being measured. Mediators are part of the complex process that ultimately produces the program effects. For example, in teacher PD intended to ultimately raise student achievement, a mediator may be the changes to classroom instruction. An inclusive study of program effects would include all of the elements of Weiss’s framework. However, the study described by this manuscript focuses on the initial phase of the model, treatment fidelity.

Fidelity as a formative evaluation. Investigating fidelity can also provide insight into the characteristics of implementation of an intervention in organizational settings (Weiss et al., 2013). When implementation is closely monitored, evaluators can gain insight into *why* a particular intervention succeeds or fails to become fully implemented (Harachi, Abbott, Catalano, Haggerty, & Fleming, 1999). For example, school leaders

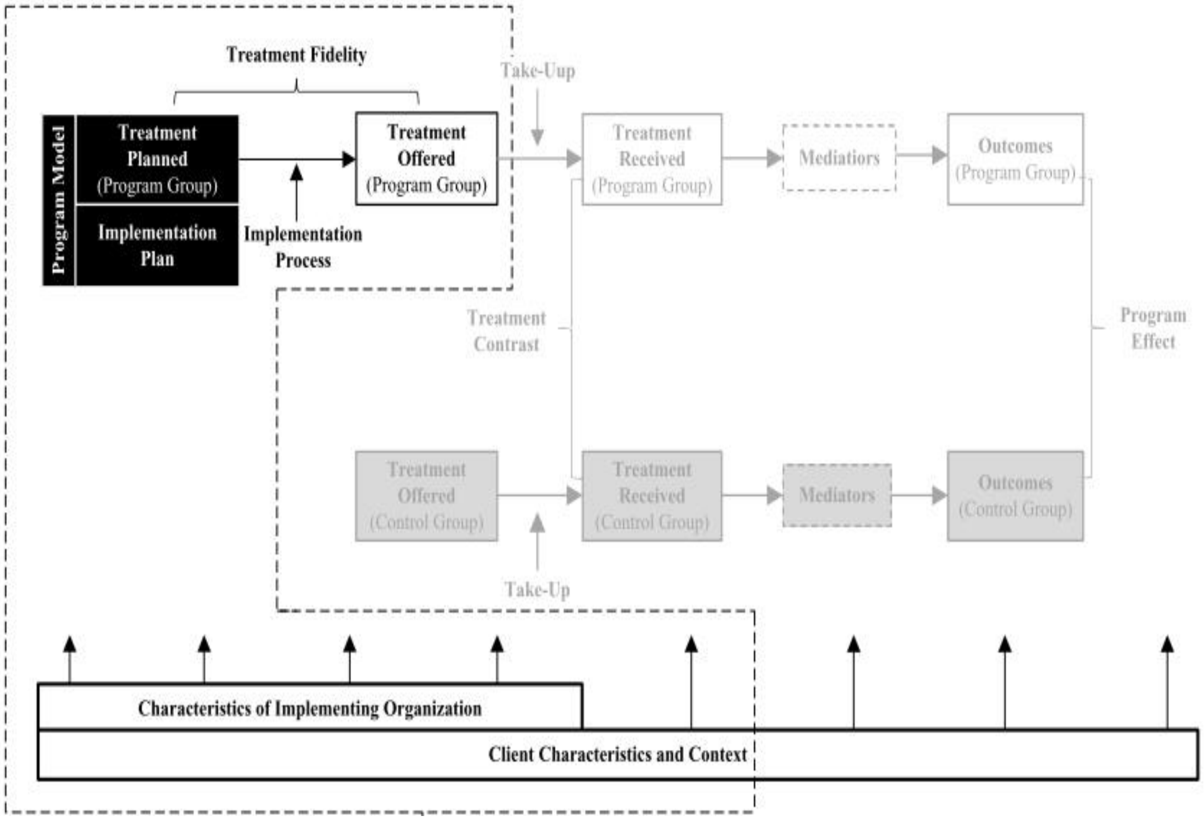

Treatment Fidelity
Take-Up
Program Model
Treatment Planned (Program Group)
Implementation Plan
Implementation Process
Treatment Offered (Program Group)
Treatment Received (Program Group)
Mediators
Outcomes (Program Group)
Treatment Contrast
Program Effect
Treatment Offered (Control Group)
Treatment Received (Control Group)
Mediators
Outcomes (Control Group)
Take-Up
Characteristics of Implementing Organization
Client Characteristics and Context

Figure 1. A framework for studying program effects that includes measurement of implementation fidelity. Taken from Weiss et al. (2013). A conceptual framework for studying the sources of variation in program effects. MDRC Working Papers on Research Methodology.

may find that the time required for daily teacher collaboration within the school day is impossible to provide. However, the evaluation may suggest that teachers provided with extended paid time are more likely to implement a program with fidelity than teachers who are not compensated for additional time commitments. Leaders can use this information to make decisions regarding resource allocation. Similarly, through early and regular measurements of implementation fidelity, leaders can provide rapid feedback to practitioners who are learning new techniques (Harn et al., 2013; Webster-Stratton, Reinke, Herman, & Newcomer, 2011). Formative feedback developed from investigations of fidelity may increase the likelihood that the intervention will be delivered as intended (Codding, Feinberg, Dunn, & Pace, 2005; Mortneson & Witt, 1998).

In addition to formative information, studying implementation also reveals how likely a program is to be implemented with high fidelity beyond initial or pilot trials. If a program is extremely difficult to implement as intended, it may not be practical or sustainable, regardless of whether the desired outcomes have been achieved (Dusenbury et al., 2003). There are often subtle components of the implementation that were influential to the success of the program that may or may not be possible to replicate (Wolery, 2011). For example, an evaluation may reveal that teachers' enthusiasm for the intervention predicted higher fidelity. However, increasing the enthusiasm of teachers with lower fidelity may prove to be a challenge.

Practitioners can also see how the implementation changes a wide range of organizational systems and behaviors, perhaps some of which were not originally targeted (Dusenbury et al., 2003). Information regarding unanticipated system changes is not only valuable to the actual implementers, but to those charged with allocating resources (Century et al., 2012). For example, school leaders looking to increase collaboration regarding student behavior may implement cross-curricular teaming structures among staff. In analyzing the intended practice, evaluators may find that curriculum-based collaboration has also increased. Accordingly, school leaders may look to support such unpredicted changes in practice through increased resource allocation.

How Is Fidelity of Implementation Measured by Evaluators?

Historically, schools have not been given consistent direction on measuring program implementation (Dusenbury et al., 2003; Harn et al., 2013). Recently, however, an increased focus on including implementation measurement in evaluation studies has forced researchers to abandon the concept of black-box approaches to program evaluation (Harachi et al., 1999; Mowbray, Holter, Teague, & Bybee, 2003; Zvoch, 2012).

Numerous guidelines for approaching fidelity investigations through measurement of *critical components* have been developed (Bond et al., 2000; Mowbray, Bybee, Holter, & Lewandowski, 2006; Mowbray et al., 2003). Hall and Hord (1987) describe critical components as the “building blocks” (p. 117) of the intervention. The building blocks are the components of the intervention that are deemed most crucial to program success. The identification of critical components underlies the process of measuring fidelity of implementation in that they allow evaluators to specify active program ingredients and uncover deviations from the intended model (Mowbray et al., 2003). Additionally, by defining and basing evaluations on critical components, evaluators can investigate whether the treatment group is actually receiving a different experience than control group participants, or if a program differs significantly across multiple sites—such as different high schools in a given district (Mowbray et al., 2003) Although other researchers use slightly different nomenclature, there is consistency in the notion of programs having specific features that must be considered when studying fidelity of implementation (Century et al., 2012).

The steps to using critical components to frame a fidelity study were summarized by Teague, Bond, and Drake (1998): (1) identify the indicators or critical components of the intervention, describing both the operational definition of the components and the methods used for measurement; (2) collect the data to measure each indicator or component; and (3) examine the data in terms of reliability and validity.

Identifying the critical components of the intervention. Mowbray et al. (2003) describe three approaches to developing fidelity criteria: (1) consult the program model of the intervention, (2) obtain expert opinion, and (3) consult the participants involved. Using the program model is the most straightforward approach, especially if the program includes key components in its manuals or other training devices (Bond et al., 2000; Christie & Alkin, 2003; Mowbray et al., 2006). Determining the critical components from the program model, however, may limit the ability to assess the intervention accurately if it has been adapted from its original design (Harn et al., 2013). For example, if a component is modified to meet the needs of a particular school culture or program, a fidelity evaluation based solely on the program model would likely indicate a lower fidelity score (Webster-Stratton et al., 2011). Flexibility within implementation, as described by Cohen (2008), suggests that adapting the original design—as in approaches (2) or (3) noted above—can have positive impacts on the intervention, and that evaluators finding lower fidelity results due to adaptations should further investigate the changes before allocating resources to increase fidelity (Harn et al., 2013; Webster-Stratton et al., 2011).

Organizing the components. The critical components of the intervention can be further described as either structural or procedural (Knoche, Sheridan, Edwards, & Osborn, 2010; Mowbray et al., 2003; O'Donnell, 2008). The structural components are those that provide the framework of the intervention, and the processes that define the way the framework is delivered (Mowbray et al., 2003). For example, structural components may include the use of required materials or the amount of time spent on a particular topic, or the contextual conditions such as student-to-teacher ratios or length of class periods (Durlak & DuPre, 2008; Harn et al., 2013). Process components tend to focus more on behaviors and interactions of teachers and students (in educational settings), or possibly doctors and nurses (in health care settings) (Century et al., 2012; O'Donnell, 2008). The organization of components by structure or process requires the researcher to document the interactions with the intervention (process) as well as the core activities themselves (structure). The distinction of components into structure and process also aids in the application of evaluations by allowing leaders to apply resources (increased training, guidance and feedback, or modifications to contextual conditions) to the components (structure or process) that are in greatest need (Durlak & DuPre, 2008; Dusenbury et al., 2003; Kaderavek & Justice, 2010).

Measuring the critical components of the intervention. Tools to assess fidelity to the critical components typically come in the form of checklists or measures that have been scaled, along with associated rubrics (Bond et al., 2000; Century et al., 2010; Mowbray et al., 2003). Ideally, these checklists or rubrics have been developed as a part of the program design, field-tested, and improved by previous users. Monitoring the application of the components can be achieved through direct observation, self-assessments by the practitioners, or a combination of both (McKenna, Flower, & Ciullo, 2014). For example, Positive Behavior Intervention and Supports (PBIS; Sugai & Horner, 2002) is a common school-wide intervention program used to improve the overall climate and culture of schools (Bradshaw, Koth, Thornton, & Leaf, 2009). Researchers at PBIS Maryland have designed a tool called the Implementation Phases Inventory (IPI; Bradshaw, Barrett, & Bloom, 2004). PBIS coaches use this tool to observe school practices to characterize the school as being at a particular level of implementation

(Bradshaw, Debnam, Koth, & Leaf, 2008). Coaches using the IPI assign a PBIS Level Rating for the school that can be used to track progress and plan further professional development. The IPI measures the critical components of each PBIS level with respect to adherence, quality, and dosage. PBIS coaches use a checklist aligned to the design as they observe teacher practices (adherence). The coaches have been trained, through PBIS, to make judgments on the quality delivery, and indicate their findings on the checklist as well. Finally, school records are used to measure how many students receive the particular components of the intervention (dose).

Challenges in measuring implementation fidelity were described by researchers at the Oregon Social Learning Center, who measured the implementation fidelity of the Oregon Model of Parent Management Training using the critical components described in the program manual (Forgatch, Patterson, & DeGarmo, 2005). Researchers found that the components could be organized into adherence and quality exclusively. The Fidelity of Implementation Checklist (FIMP; Knutson, Forgatch, & Rains, 2003) was used to measure the components. The need for flexibility by practitioners and the varying degree of client engagement became a challenge when applying the binary checklist. The FIMP consisted of direct observations and video recordings of sessions (Forgatch et al., 2005). The primary goals of the evaluation were to identify the psychometric properties of the FIMP and to measure the efficacy of the training. A Cronbach’s alpha reliability analysis of the raters revealed a range of 0.87–0.95, depending on the component. The correlation between the items ranged from 0.71 to 0.90. The evaluation revealed that fidelity of implementation could be shown to account for 30% of the change in the parental behavior. In addition, the researchers found that practitioners used their professional experience to adapt the components to meet the needs of the individual recipients. In doing so, the level of fidelity was lowered, although the change may have been warranted. These researchers recommended that observers record and review videotapes of sessions in order to code all activities (Forgatch et al., 2005). The studies presented above give insight to the opportunities and challenges of measuring the fidelity of implementation within a program evaluation. The studies also present methods to limit the impact of challenges when designing a program evaluation. For example, the use of simple checklists causes judgments to be made too narrowly. Preferably, comprehensive descriptions of components with progressive rubrics should be used when available.

The specifics of conducting observations present additional challenges. For instance, the timing of the observations may affect the results (Bond et al., 2000; Yeaton & Sechrest, 1981). Studies have shown that fidelity to program adherence can vary over time (Dusenbury et al., 2003; Zvoch, 2009). Therefore, repeated measures of fidelity over time are preferable to a single-point data collection (Zvoch, 2009). Multiple measures yield a better understanding of the average adherence when implementation is likely to vary over time. Additionally, the rate of change of implementation may be determined. With multiple measurements over time, it is possible to examine whether implementation increases, decreases, or remains unchanged over the course of a school year.

The general feasibility of fidelity measurement may also impact the evaluation of program implementation (Mowbray et al., 2003). At times, fidelity measures are aligned to components that can be practically measured; however, they do not accurately reflect the scope of the intervention (McGrew, Bond, Dietzen, & Salyers, 1994). Thus, results of these studies that do not address every component in the intervention are limited to the components that are measured. For example, a component of an intervention may be focused on student academic talk. A practical measure of student talk is to record the ratio of teacher talk to student talk, or even more simply, the number of minutes per class that a student is talking. Although measuring the quantity of student talk is straightforward, the measurement would not describe whether the talk was academic or not. By not measuring the academic nature of the talk, the scope of the component would not be fully assessed.

In order to measure student talk more completely, observers would need to measure the quantity of talk as well as the substance of the talk. Observers need to be in much closer proximity to students to do this, which may cause students to change their behaviors, or, at least, increase the difficulty of observing a wide range of students. Alternatively, audio recordings could be obtained, transcribed, and coded into varying degrees of academic talk. In classrooms where the student talk is directed to the instructor and from single students at a time, the use of recordings may be practical. However, in classrooms where student talk is directed to each other in dyads or small groups, a practice that is considered beneficial (Bickmore & Parker, 2014), numerous recording stations would need to be set up at multiple points in the classroom. The equipment demands of setup and the personnel demands of transcribing and coding over multiple classrooms may render the approach impractical.

Determining the reliability and validity of the measures. Data collected must first be analyzed for reliability and validity prior to making meaningful conclusions (Mowbray et al., 2003). Reliability generally refers to the ability of a test or other technique to yield consistent results (Babbie, 2007). There are two forms of reliability particularly relevant to this case study. Reliability between observers, or inter-rater reliability/consistency, is important if more than one person will be making observations. Secondly, the reliability between the scores obtained from the different items in the instrument should agree with one another. For example, it is relevant to verify that scores on items that represent a particular construct positively correlate with one another.

Reliability indices should first account for the level of agreement on the judgments of the same event. The simplest measurement is in the form of a percent agreement. Percent agreement, however, is not considered to be adequate, as it does not take into account the agreement that would be expected due to chance (Hoehler, 2000). Cohen’s kappa is a simple extension of the rate of agreement that corrects for the agreement expected by chance. The kappa statistic is designed for use with nominal or ordinal data, preferably when only binary judgments are made (Morgan, Leech, Gloeckner, & Barrett, 2013). Although the kappa statistic was designed for binary scales, it is often applied to graded measurements due to its relative ease in calculation and interpretation (Morgan et al., 2013).

A second approach is to account for the internal consistency of the item responses by using Cronbach’s alpha (Bond et al., 2000). Internal consistency refers to the agreements among items on a particular measure that evaluate a specific construct. In an evaluation of teacher practices, observations may be made using a particular rubric that evaluates a standard or domain. The rubric for a particular standard may include multiple indicators. The groups of indicators for a particular standard should yield a similar result, regardless of the observer. The measurement of internal consistency using Cronbach’s alpha usually utilizes three steps. The first is the determination of the alpha itself, providing evaluators an indication of the agreement of scores on the items. Next, an analysis of the inter-item correlations is made, allowing evaluators to determine which scores agree with or contradict each other. Finally, the alpha is repeatedly measured by removing single items one at a time. The alpha with an item removed can be compared to the alpha with all items included. Alphas that are increased when particular items are deleted suggest a particular item is lowering the internal consistency and should be considered for removal from analysis (Morgan et al., 2013).

Determining validity of measures. Validity refers to the degree to which the data support the adequacy and appropriateness of the interpretations and actions that they derive (Messick, 1994). In quantitative studies, three forms of validity should be considered: content validity, predictive (concurrent) validity, and construct validity (Creswell, 2014).

Content validity is the ability of a test to measure the content it was intended to measure. Content validity ensures that the measure adequately captures the breadth of the target. Content validity can be measured using field experts to review items, review descriptions of the content, and make judgments as to the completeness of the measure (Polit & Beck, 2006).

Predictive, or concurrent, validity is the degree to which scores predict or correlate with other measures of the same content or construct (Creswell, 2014). For example, both the College Board's ACT and the National Assessment of Educational Progress (NAEP) include sections designed to measure students' "reading ability." High concurrent validity between the ACT and NAEP would indicate that students scoring high in the reading section of the ACT would also score high in the reading section of the

NAEP. Predictive validity indicates whether or not a measure adequately predicts a criterion. An example would be if the College Board's ACT exam accurately predicts future college success (Babbie, 2007). Predictive validity can be measured using regression analysis or similar inferential statistics (Morisky, Green, & Levine, 1986).

Related to content validity is the concept of construct validity. The term *construct* refers to abstract or difficult-to-observe properties, such as motivation or personality, as opposed to easy-to-define observables like pH and age (Thorndike & Throndike-Christ, 2011). Construct validity refers to the degree to which a study accurately measures the intended construct (Tindal & Marsden 1996). Messick (1994) describes two general threats to construct validity: "construct underrepresentation" occurs when a measure is too narrow to fully describe the construct, whereas "construct-irrelevant variance" arises when the measure is too broad and includes indicators aligned to other constructs.

The threats to validity should be addressed when designing a program evaluation. In order to limit the threats to measurement validity, evaluators first need to thoroughly understand the components of the intervention. Understanding can be derived from qualitative data on the specific pieces of the implementation process through the involvement of the people closely involved with the intervention (Brunette et al., 2008; Singh & Fletcher, 2014). A complete understanding of the components should be developed through a review of the program model, but also through the involvement of key stakeholders and experts (Brandon, 1998; Mowbray et al., 2006). Brandon (1994) synthesized the findings of four studies to develop guidelines for including stakeholder and expert input, in addition to a review of the program model for the purposes of limiting threats to measurement validity. He concluded that researchers should ensure that the groups included have the appropriate experience and able to participate. They should also take care in developing thorough methods for gathering stakeholder feedback. Finally, stakeholder groups should have equitable participation in the feedback processes, meaning simply that “no stakeholder group’s expertise is ignored in the evaluation and decisions making process” (Brandon, 1998 p.8).

The use of an instrument, ideally a graded rubric, is the central element of fidelity evaluations. Therefore, the ability for the instrument to generate reliable and valid data is paramount to the confidence that underlies analysis. The difficulties in obtaining reliable and valid observational data are highlighted throughout the current study.

The Investigation of a Specific Intervention

Constructing Meaning (CM) training is a product by E. L. Achieve, an educational consulting company. The basic premise of the program is that English Language Development needs to be integrated throughout all curricula, not just in an *English* class (Dutro, 2009). Requiring all teachers to use strategies in language and literacy development is a shift in pedagogy, especially at the secondary level. Secondary schools are typically segmented into distinct subjects, where the science teacher is responsible for the science content and the language arts teacher is considered solely responsible for literacy development (O’Brien, Stewart, & Moje, 1995). The students, as well as the teachers, may realize this segmentation. Measor (1984) found that students’ actions and behaviors varied significantly throughout the day, depending on their perceptions of the current course. For example, students were more likely to make language convention errors in a science class than a language arts class, where they perceived the practices to be more relevant. In order to shift the perception that language convention is irrelevant in non–language arts classes, CM training provides strategies for teachers to utilize within their content areas to improve the overall academic language proficiency of their students.

CM training is designed to enable teachers to lead students to develop their English language proficiency while still meeting the rigorous demands of content area courses. The foundational basis of CM includes procedures for the following:

- Ensuring both a *content* and *language* objective for every lesson.
- Using a *functional* language approach to instruction. Typical language functions include comparing two ideas, persuading an audience, or defending a claim.
- Dividing introductory lessons into discrete chunks to scaffold students toward longer, more complex activities.
- Explicitly teaching language with opportunities for written and oral practice in every course of study.

CM teacher training is provided in a three-day seminar where teachers learn background research, teaching strategies, and methods for adapting existing lessons. Following the training, teachers are provided with institutional handbooks as well as access to instructional coaches for support. The training begins with a background of relevant concepts in language development. Teachers then transition to learning specific strategies to be implemented in their classrooms. Strategies include the use of language targets, the use of sentence frames, and tools to scaffold the "bricks and mortar" of their lessons (Dutro & Moran, 2003). For the purposes of CM, *bricks* refer to the vocabulary that is specific to the course of study. As an example, the terms *stoichiometry*, *amphoteric*, and *monoprotic* would all be considered "bricks" of a high-school chemistry course. The “mortar” are academic words that are consistently used regardless of content, such as *therefore*, *analyze*, or *however*.

The strategies taught to the teachers extend through each of the foundational principles described above. Following the strategies portion of the training, teachers are given time to adapt their curricula (casually referred to as “CMing”). Teachers are taught to adapt their curricula by applying the strategies they learned to meet the overall goals described by the critical components. Finally, teachers present mock lessons and self-evaluate their work based on a rubric developed in alignment with the critical components.

Critical components of constructing meaning. E. L. Achieve, the developer of the CM training, has designated five areas as critical components:

(1) *Understanding Backward Design.* This includes designing instruction that addresses the cognitive and linguistic demands required to meet stated student learning goals.

(2) *Language as a Part of Content Teaching.* This component requires creating opportunities to learn both content “bricks” and functional “mortar” throughout instruction.

(3) *Oral Language Practice.* This refers to instructional designs that allow for structured peer interaction for students to use the target language (English) of the learning goal, including students who may have limited English language proficiency.

(4) *Interactive Reading and Note-Taking.* This describes the use of comprehensive strategies and note-taking tools to facilitate the navigation of complex text and increase student independence.

(5) *Academic Writing Support.* This final component prompts teachers to provide tools and facilitate processes that support students in producing complex academic writing.

Each of the critical components align to one or more of the fundamental concepts and are operationalized by teachers using specific strategies presented in the training in their classroom practices.

As described above, one fundamental concept within CM practices is the *explicit teaching* of academic language through the strategies delivered in the *Language as a Part of Content Teaching* critical component. Explicitly teaching language involves direct and unambiguous strategies to teach academic language acquisition (Rosenshine, 1987). Criteria for qualifying a specific strategy as *explicit* were summarized by Archer and Hughes (2011) and are in line with the strategies presented in CM training. In a meta-analysis of 49 experimental and quasi-experimental studies, researchers investigated the effect size of various approaches to second-language instruction on student achievement (Norris & Ortega, 2000). The approaches were classified into four categories: implicit and explicit instruction using the Focus on Form (ForM) approach, and implicit and explicit using the Focus on Forms (ForMS) approach. The ForM approach teaches the forms of language as they come up incidentally in a student’s academic conversation. In contrast, the ForMS approach teaches linguistic elements in discrete lessons (Sheen, 2002). The assessments used varied through the experiments in the meta-analysis, but were grouped into four categories: metalinguistic judgments, selected response, constrained constructed response, and free constructed response. Overall, the researchers found that on various student performance outcomes, explicit language instruction had a mean effect size over one half of a standard deviation greater than that associated with implicit instruction. These results suggest that utilizing explicit strategies, such as those presented in CM training, may lead to relatively stronger English Language Arts (ELA) achievement.

As described above, an additional premise for consideration is the explicit teaching of academic English language throughout the curriculum, not just in literacy courses (E. L. Achieve, 2014). Teaching language within other content courses has been shown to increase the contextualization experienced by students and, in turn, increase levels of achievement (Tompkins, Campbell, Green, & Smith, 2014). Additionally, providing professional development in academic language development to all teachers serves as a more pragmatic approach in light of the current standards. In both the Common Core State Standards for Mathematics (CCSS; National Governors Association, 2010) and the Next Generation of Science Standards (NGSS; Lead States, 2013), standards include requirements for communication, collaboration, and text complexity. The added requirements strengthen the case for language acquisition to be taught in all classes, resulting in a need for professional development opportunities involving all teachers (Archer & Hughes, 2011) such as CM.

The design of lessons through the use of strategies delivered in component (1), *Understanding Backwards Design,* builds on the concept that language instruction should occur in all content classes. Teachers are instructed to begin the design of the lesson with both content and language objectives, giving students language goals in addition to content objectives (Ferretti, MacArthur, & Dowdy, 2000). The use of clear objectives allows teachers and students to navigate the different standards that are directing the class. For example, a typical high school biology course can be aligned to Common Core Literacy Standards, NGSS, and locally adopted standards for English Language Learners (ELL) curriculum (Valdés, Kibler, & Walqui, 2014). Clear short-term objectives allow students to understand the outcomes they are expected to achieve by completing their daily assignments. For example, in the current study, the district has aligned every course to "learning targets." Learning targets serve as the classroom-level guide for the implementation of broad standards, and allow teachers to appropriately design their instruction to ensure alignment. Specifically, teachers are able to explicitly express their high expectations for students, ELLs in particular, who may have experienced lower expectations in other school settings (Echevarria, Frey, & Fisher, 2015). Standards, and associated learning targets, set the benchmarks for students as they progress through the school system.

Additionally, including language objectives supports the concept of language instruction across the curriculum, as described above (Vacca & Vacca, 1989). Including a language objective in a content class is a method to teach language explicitly. Norris & Ortega (2000) completed a comprehensive meta-analysis comparing instruction to student outcomes in writing. The dependent variable was described as students' demonstration of language. The nature of the meta-analysis did not allow for a single common measure to be used; however, the measures across the study were coded into four groups: metalinguistic judgments, selected response, constrained constructed response, and free constructed response. The parameters of the meta-analysis defined the independent variable as literacy instruction being implicit in nature. The experimental group contained only classes where explicit language instruction was used. Explicit instruction was defined by DeKeyser (1995) as instruction that requires students to attend to specific linguistic rules or forms. For example, two science teachers may be using the same article related to the mechanisms of photosynthesis and respiration. One teacher may ask students to identify the literary moves that the author makes in comparing photosynthesis to cellular respiration. The other teacher may restrict the students' tasks to simply content-specific comprehension, such as understanding the different roles of energy in the two processes.

Researchers found an average effect size of 0.75 throughout the studies, with a pre-test and post-test measuring the impact of direct language instruction as described above. These outcomes suggest that explicit language instruction may lead to higher outcomes across subject areas. It should be noted that there are multiple strategies to instruct language directly, and that developing and displaying a language target cannot be considered complete language instruction. However, the backward design of lessons and units from a defined language objective is critical to the CM approach (Dutro, 2009). The actual design of the lesson will ideally include strategies from all of the remaining components of *Interactive Reading and Note-Taking*, *Academic Writing Support*, and *Oral Language Practice*.

Interactive Reading and Note-Taking summarizes strategies intended for the production of work using academic language derived from content area texts, lectures, and other learning opportunities. Teachers can provide discrete scaffolding to more complex objectives through interactive reading and note-taking. Providing such scaffolding allows students to participate in more conceptually abstract activities than they would otherwise be able (Lucero, 2013). Specific strategies for the interaction with notes, as opposed to allowing students to take notes passively, was shown to have a modest effect size of 0.22 on students' post-test performance from a meta-analysis of 57 studies comparing note-taking to non-note-taking strategies (Kobayashi, 2005). Additionally, a key piece of *Interactive Reading and Note-Taking* is the summarization of key learning. According to a meta-analysis presented to the Carnegie Foundation, there is an effect size of 0.82 on assessment of "quality writing" when students are explicitly taught to summarize texts (Graham & Perin, 2007). The CM participant manual offers more than 15 distinct strategies for teachers to use in order to increase the interaction of students and their reading or note-taking assignments.

Academic Writing Support provides strategies to shelter the challenges of language acquisition away from content knowledge. The approach stems from the theory that knowledge is transferable between languages (that is, if you understand something in one language, you understand it in the other) (Bangert-Drowns, Hurley, & Wilkinson, 2004). Often, students struggle with representing their knowledge in a second language (a language challenge), and this is misrepresented as a content challenge. By using strategies such as sentence frames and instruction specifically targeted to vocabulary instruction, teachers can help students communicate their learning clearly even as language development is still occurring (Bangert-Drowns et al., 2004; Graham & Perin, 2007). In a meta-analysis of 123 studies, researchers were able to identify specific areas of writing instruction and summarize their effect sizes on post-test performance. Of the areas identified, utilizing explicit instruction to teach students the components of writing, such as pre-writing, drafting, and revising, yielded an average effect size of 0.82 relative to “writing quality” across the studies that were analyzed (Graham & Perin, 2007).

Opportunity to increase student achievement. The middle school that is the site of this study has consistently underperformed on standardized tests, particularly in ELA. In the most recent state report card, the school earned a Level 3, placing it in between the 15th and 44th percentile of all middle schools in the state. Student achievement data that is disaggregated by subgroup indicates a predictable achievement gap. Approximately 20% fewer Hispanic and ELL students in the school meet state benchmarks in ELA. Recently, school and district leaders have committed to supporting teachers in improving students’ outcomes through the use of high-quality PD. The focus of the professional development has been primarily around academic language instruction.

Recently, Sheltered Instructional Observation Protocol (SIOP)—which is used for observing teachers who are using specific instructional strategies that target the development of academic English language—was used to increase teachers’ understanding of best practices in language development. The protocol is arranged around eight areas, each of which can be observed during classroom instruction. Teachers observed using a high degree of fidelity to these eight areas receive a high SIOP score. Use of the SIOP as a tool to measure implementation fidelity was studied in a large urban school district (Echevarria et al., 2011). Overall, researchers found that the greater the SIOP score, the greater the student achievement, with the SIOP score explaining approximately 21% of the variance in student achievement. Despite the claims that SIOP practices can raise student achievement, the middle school of this case study has shifted its focus. According to the school principal, the teachers were supportive of SIOP, but they were looking for more specific strategies than those provided. As a replacement, Constructing Meaning training was chosen because district leaders felt the approach of CM, including the fundamental basis and critical components described above, would serve as a follow-up to SIOP and provide continued support to teachers as they explicitly teach language acquisition in their classrooms.

Using a Qualitative Approach to Fidelity Studies

Although much of the fidelity research has been quantitative, studies that are designed to understand processes and events, such as program implementation, may benefit from including a qualitative approach (Maxwell, 2013). For example, researchers at Dartmouth Medical School conducted a follow-up qualitative study to further examine quantitative implementation data (based on observational checklists) derived from the provider's use of a mental health intervention protocol (Brunette et al., 2008). The researchers used field observations and semi-structured interviews to understand the "facilitators and hindrances" of the specific implementation. The results yielded meaningful claims around both a priori and unpredicted characteristics of implementation. The evaluators were able to organize the hindrances that they uncovered into specific themes of leadership, supervision, staff turnover, consulting with experts, and finances. The evaluators were also able to provide recommendations to hospital management based on each of the themes. For example, one recommendation regarding the theme of leadership was to ensure that the staff understood the level of prioritization the intervention had compared to other hospital objectives. Sites that demonstrated high levels of fidelity were able to clearly prioritize the intervention through policy, financial, and human resource decisions. Participants in sites with low fidelity felt that their leader or leaders had failed to clearly establish the interventions as a priority. Evaluators presented leadership with a recommendation to take steps to clearly show the intervention as a priority. The leaders' actions on the recommendations included changes in personnel, the development of policy, and increases in communication to the staff from the management.

Similarly, researchers at the United Kingdom's National Institute for Health Research investigated changes to the behaviors of both practitioners and recipients using an implementation fidelity framework that included qualitative methods (Dyas, Togher, & Siriwardena, 2014). The researchers designed interview questions to investigate both adherence to the model and participant responsiveness. The interview responses allowed researchers to gain a better understanding of the pilot data and to better explain the quantitative data. Specifically, researchers were able to ask participants questions directly related to the quantitative data and report on their responses. The combination of quantitative and qualitative data allowed the evaluators to make more specific recommendations to leadership—in particular, in areas in need of improvement.

The combination of quantitative and qualitative data in an evaluation of fidelity may be particularly useful when the purpose for the study is of a formative nature. Quantitative studies, with strong experimental designs, are well suited to describe cause and effect relationships; they are not as well suited to questions of a "how" or "why" nature (Collins, Onwuegbuzie, & Sutton, 2006). Qualitative methods, such as interviews, surveys, and focus groups gather information about the human experiences of the program being evaluated. The descriptions of the experiences can yield information on the context-specific beliefs and biases that contribute to the level of implementation of the program being evaluated (Sankar, Golin, Simoni, Luborsky, & Pearson, 2006). In addition, the benefits include the opportunities to hear the perspectives of the providers as to what components of the intervention are presenting challenges for implementation. Understanding the perspectives of the providers would not be apparent in the quantitative data alone.

Mixed-methods research attempts to combine the strengths of quantitative and qualitative methods into a single design (Babbie, 2007; Johnson & Onwuegbuzie, 2004). Mixed-methods research allows researchers to obtain a more complete understanding of the phenomena they are studying rather than using a single method (Hesse-Biber & Johnson, 2013; Johnson & Onwuegbuzie, 2004). Quantitative analysis tends to be very objective and maintain a value-neutral stance in the discussion. Conversely, studies that are solely qualitative utilize subjective analysis and can include a value-specific approach in the discussion (Tashakkori & Teddlie, 1998). Including both quantitative and qualitative analysis can be used to explain or interpret initial findings, explore an observed phenomenon, or address a question from multiple levels.

A Mixed-Methods Approach to the Fidelity of Implementation of Constructing Meaning Training

In line with other program evaluation studies of implementation fidelity, a formative evaluation of CM implementation was conducted at a middle school in a Northwest Oregon School District[1]. The concept of treatment fidelity was used to design a study that

[1] "Northwest Oregon School District" is being used as a pseudonym to ensure confidentiality.

measured the implementation of the critical components of Constructing Meaning training. The measurement of the implementation of the components resulted primarily from classroom observations that utilized the *Refining Our Practices Rubric*. The observational data was analyzed using primarily descriptive statistics. The results of the quantitative data analysis were presented to teachers during the qualitative phase of the study, along with survey and interview questions, in order to understand their perspectives on the quantitative findings. In line with applied research, this study addressed a specific school and district need by conducting a comprehensive evaluation of the implementation of CM practices. Neither the middle school, nor the larger district, had a systematic evaluation plan in place. The case study described here was used to determine the level of fidelity of CM training, understand the perceptions of the providers (teachers), and make recommendations regarding implementation of CM practices. My conclusions and recommendations result from investigation of the following research questions:

- RQ1. How successfully has the faculty of a local middle school implemented the critical components of Constructing Meaning training?
    - To what degree have the critical components been implemented?
    - Is the variation in implementation predictable?
    - How does the degree of implementation compare to a determined threshold?
- RQ2. What are the conditions that favor or hinder a high degree of implementation fidelity in Constructing Meaning practices?

CHAPTER III

METHODS

A mixed-methods design was used to investigate the research questions presented in Chapter II. The following chapter describes the research setting, participants, measures, and analysis procedures.

Setting and Participants

The Northwest Oregon School District, where the middle school of this case study is located, has adopted Constructing Meaning (CM) as a major source of professional development, specifically at the secondary level. The district has communicated a commitment of having every middle school teacher trained in CM within the next three years. The majority of the middle school teachers have yet to be trained, and the district must make a significant resource allocation in order to meet the goal. With school budgets still below pre-recession levels, allocation of resources is closely scrutinized and school leaders must continually monitor the return on investment (ROI) in programs and practices. As a result, the district agreed to participate in this study as a formative implementation evaluation in a pilot school that has been involved with CM training for the past three years.

This study took place exclusively within one middle school in the Northwest Oregon School District. The district is one of the largest in the state, serving approximately 40,000 students. The school's demographic composition is approximately 42% white, 36% Hispanic, 9% Asian, 5% black, 1% Pacific Islander, 1% Native American, and 6% multiracial students. Approximately 37% of the students are English Language Learners (ELLs), 16% receive special education services, and 64% participate in the Federal Free and Reduced Meals program. The middle school includes grades six through eight and is considered a comprehensive middle school without a specialty program (such as the International Baccalaureate's Middle Years Program).

The study participants included middle school teachers and students. The school employs 52 certified teachers, 34 of whom have been trained in CM practices. Four of the trained teachers opted out of the study. The participating teachers (*n* = 30) include 10 Math, 8 Science, 4 Humanities (combined Language Arts and Social Studies), 3 Special Education, 2 Art, 2 ESL, and 1 Physical Education teacher. The teachers varied in teaching experience from 1 to 25 years, with a mean experience of 9.85 years (SD = 5.02). Forty-five students, 15 at each grade level, had a direct role in the study by participating in focus groups. The students were selected at random from grade level lists and were given the option to participate. All of the students agreed to participate and their guardians granted permission. However, for each group, some students were absent on the day of their assigned focus group, presumably due to illness. The resulting groups consisted of 13 sixth graders, 14 seventh graders, and 11 eighth graders. Twenty were male and eighteen were female. Forty-one percent were designated as English Language Learners (either active or monitored). The students were organized into eight focus groups of four to five students each. The groups were set in order to minimize the disruption to the students' school day. Students were pulled from elective or teaching assistant periods, when possible. The grade level remained constant with each group and the male/female ratio was as even as possible.

The district does not have a uniform model for ELL inclusion throughout its schools. Some schools opt for a "pull out" model, where specific language instruction is delivered in a class that is distinct and not connected to the grade level language arts class. In contrast, other schools, including the study site, opt for a more inclusive model where all ELLs continue with grade-level Language Arts and Social Studies classes. As a result, individual classes are a heterogeneous mix of students, closely reflecting the overall demographic of the school. All of the teachers in the study had classroom ELL populations between 29% and 45% of the total student makeup.

Constructing Meaning training has been a significant source of professional development at this middle school during the past three years (Brock, personal communication, September 15, 2014). As described in Chapter II, the training was selected as a follow-up or continuation of previous work to implement Sheltered Instruction Observation Protocol (SIOP) techniques from trainings that occurred from approximately 2006–2009. The school employs two “instructional guides” that have been certified by E. L. Achieve as CM trainers available for additional training and support. The instructional guides earned this certification through a “train the trainers” process facilitated by E. L. Achieve. The process to become a certified E. L. Achieve trainer requires participation in two additional workshops beyond the initial training. The first additional workshop is called a +2, referring to the two days spent reviewing videotaped examples of implemented CM practices, training on observations, and discussions of quality feedback. The second additional workshop is called “District Leadership” and includes shadowing other trainers, review of local achievement data, training in E. L. Achieve’s approach and practices, and implementation planning.

The guides were essentially “on-call” to observe teachers and provide feedback, help develop curricula, or assist in the delivery of lessons. The use of instructional guides by the teachers was voluntary, and the frequency of use was not recorded. However, casual conversation with the guides revealed that they felt they were frequently utilized by some teachers and rarely accessed by others.

Research Design

The purpose of this mixed-methods study was to monitor and describe the implementation of CM practices at the study site. Therefore, this study investigated the experiences of teachers and students within the school that has piloted the training in order to gain insight into the nuances of implementation that the district could use for future planning. The insight provided would be framed around the success of implementation and the conditions that favored or hindered implementation.

Phases of research. There are a variety of design approaches within the field of mixed-methods research. Ivankara (2006) identified more than 40 different mixed-methods research designs referenced in the literature. However, Cresswell et al. (2003) describe the six most commonly used designs. Within the six designs, three are concurrent—where the quantitative data collections and analysis occurs simultaneous to the qualitative—and three take place sequentially in two distinct phases. Researchers use concurrent designs when the goals of their studies include comparing or consolidating the quantitative and qualitative findings. Alternatively, researchers use sequential designs when the goal of the second phase is to explain or elaborate on the findings of the first phase. In the explanatory sequential design illustrated in figure 2, researchers apply quantitative methods first, followed by qualitative methods in order to understand more fully the initial quantitative findings Creswell, 2014). The current study utilized this design to investigate the two research questions (RQ1 and RQ2) presented in Chapter II.

The use of explanatory sequential design was appropriate in the current study because the components of the quantitative data had been pre-established, eliminating the need to explore the components through qualitative measures first, as would be the case in other mixed-methods designs. RQ1 was written to be investigated using primarily quantitative techniques while RQ2 was written to be investigated using primarily qualitative techniques. Beginning with quantitative data was advantageous as it provided a foundation for the qualitative measures, particularly the semi-structured interviews.

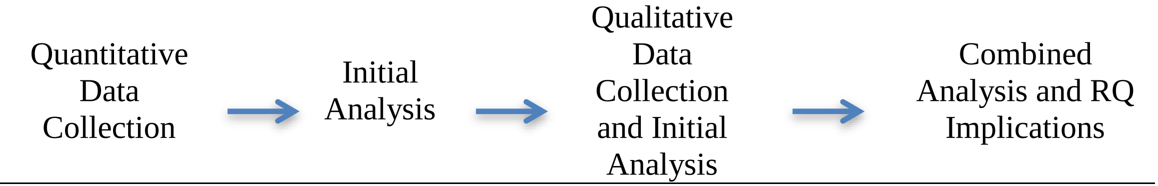

Quantitative Data Collection
Initial Analysis
Qualitative Data Collection and Initial Analysis
Combined Analysis and RQ Implications

Figure 2. Explanatory sequential design.

Success of implementation. Following the explanatory sequential model, phase one of this study included quantitative methods focused on addressing RQ1. RQ1 investigated the success of the faculty at the local middle school in implementing the critical components of CM training. Prior to this study, neither the school, nor its parent district, had set clear expectations for the level of implementation expected. Due to the lack of a predetermined standard, this study utilized personal communication with district leaders and CM trainers, a review of the CM program manual, and a review of relevant literature to develop a standard of success. The resulting standard included determining the degree of implementation of the components of CM training, the level of implementation variability between teachers, and the comparison of actual implementation to a predetermined standard. The standard is described later in this chapter.

Critical components of CM training. As discussed earlier, researchers have suggested framing implementation studies around the components of the interventions that are most critical for an acceptable implementation. The investigation in this study was accomplished through the identification, measurement, and interpretation of the implementation of the critical components of CM training. By utilizing a critical components approach, the operational definition of implementation fidelity for this study was the degree to which the critical components of CM practices were implemented by teachers at the middle school. The critical components of CM training, as defined by E .L. Achieve, are: (1) *Understanding Backward Design*, (2) *Language as a Part of Content Teaching*, (3) *Oral Language Practice*, (4) *Interactive Reading and Note-Taking*, and (5) *Academic Writing Support* (see the discussion of each of these components in Chapter II). These components were identified by the developers of CM through the review of relevant literature, the opinions of experts in the field, and follow-up dialog with participating schools across the country (E. L. Achieve, 2014). As described in Chapter II, the critical components are simply an organization of the instructional practices that most closely align with the key research-based principles of CM. According to personal communication with representatives from E. L. Achieve, past and future revisions to the critical components focus on the descriptive terminology and the specific groupings of strategies. For example, in an upcoming version of the rubric (as of October 2015), components (4) and (5) have been reworded to *Language for Reading Comprehension* and *Language for Writing Comprehension*. The goal of rewriting the rubrics is to further define the specific parameters of the critical components.

E. L. Achieve also publishes the *Refining Our Practices Rubric*, a tool that describes the adherence, quality, and to some extent the dose of each component outlined above. The application of the rubric, by trained observers, to this case study provided the basis for collection of quantitative data on implementation fidelity. The rubric is further discussed in the Instruments section of this chapter.

Variation in implementation by predictor variable. Years of teaching experience, teachers' primary subject area, and time since receiving CM training were used as predictor variables in data analysis. Although other factors are likely related to CM implementation, such as teachers' initial buy-in and previous quality of teaching, they are difficult to measure and extend beyond the scope of the current study.

Level of implementation compared to the literature. The school, the district, and the E. L. Achieve organization expect that the use of CM practices will have a positive impact on student achievement in both classroom-based and state and national standardized exams. Generally, high fidelity of implementation has been shown to increase intended outcomes (Benner, Nelson, Stage, & Ralston, 2011). However, the current study focused exclusively on the level of implementation by teachers and did not analyze student achievement data. Because of the omission of student achievement data, success was not measured by an exam achievement standard, but rather by an implementation standard. As previously described, neither the school nor district had established expectations for the level of implementation. A review of the CM program manual provided little insight as to the level of implementation that could be expected (E. L. Achieve, 2014). The references to expected timeline to achieve full implementation are vague, indicating that teachers need to “practice in the classroom to improve” and “use the Refining Our Practices” rubric as a formative tool to progress (p. 63).

The literature also does not provide a universal standard level of implementation needed to achieve anticipated results. The tolerance of limited implementation varies by intervention (Kaderavek & Justice, 2010). The use of typical standard setting models, such as the Angoff or Ebel, require substantial training and time resources not available in this study (Cizek & Bunch, 2007). Therefore this study relied on face validity to develop a success threshold. School leaders, CM instructional coaches, and the primary investigator met to discuss the intended levels of implementation as part of this study. In addition, a review of the terminology of the rubric was used to formulate a success standard. Level 2 scores for rubric items included terms that had a negative connotation, such as “rarely,” “occasionally,” “to individual students,” and “not addressed.” In contrast, the level 3 descriptions included more positive terminology, such as “frequently,” “used by most students,” “including both bricks and mortar vocabulary words.” The determination was made that all rubric items evaluating critical components should be scored at level 3 or higher when observing a “successful CM Teacher.” The expectation of “all [level] 4s, all the time” was not practical, in the opinion of the group. Accepting the “all [level] 3s” consensus meant that an implementation rate of 75% would equal “successful implementation.” Additionally, as this evaluation was intended to be formative in nature, school leaders cautioned the principal investigator to frame the 75% threshold solely as a marker for this study and not an administrative directive.

Phase two, the investigation of RQ2. Phase two utilized qualitative methods to investigate the conditions that favored or hindered the implementation of CM practices.

Qualitative research methods provide tools for achieving goals related to interpretation and understanding of social phenomena (Merriam, 2008; Maxwell, 2013; Creswell, 2014). Qualitative research has certain characteristics (Creswell, 2014). The characteristic of *natural setting* occurs when the research is conducted where participants experience the topic being investigated. In the current study, all data was conducted exclusively within the middle school where CM practices were being implemented. The characteristic of *inductive data analysis* is found within the concept of explanatory sequential design, where emerging data patterns and themes are directly investigated by specific questions or other qualitative methods. In the current study, both the initial analysis of quantitative data and the emerging themes from initial qualitative analysis were explicitly discussed in open-ended interviews. Another characteristic present in the current study is referred to as *participants' meanings.* Participants' meanings direct researchers to keep the focus of analysis on perceptions that the participants hold in regard to the issue, not what the researcher expects or desires. In the current study, techniques such as the display of negative information and member checking were used to ensure that participant meaning was included.

The qualitative methods included surveys (both closed and open-ended questions), semi-structured interviews with teachers, and student focus groups. Surveys are used in qualitative studies to describe, compare, and explain individual and organizational knowledge (Fink, 2013). The surveys in this study asked questions to solicit the teachers' perceptions of the overall training as well as the individual components. Interviews, particularly less structured interviews, allow the researcher to do more active inquiry by asking questions specific to the emerging themes of study

(Babbie, 2007; Warren, 2002). The interviews in the current study included the presentation of initial data from classroom observations as well as the themes that emerged from the structured surveys. Similarly, focus groups allow participants to provide additional details relevant to each other’s comments (Sankar et al., 2006). Focus group participants, particularly minors, may also be more comfortable in a group setting opposed to individual interviews (Ouimet, Bunnage, Carini, Kuh, & Kennedy, 2004).

Time element. This case study was completed during the 2014–2015 school year by conducting teacher observations, teacher reflections, teacher surveys, teacher interviews, and student focus groups. Teacher observations, reflections, and surveys occurred during a six-week period (February to mid-March 2015) at the beginning of the second semester of the school year. Conducting the research at the beginning of the second semester was advantageous for two reasons. Teachers who had been trained in the beginning of the school year needed sufficient time to apply what they had learned in the training. Also, by completing the observations in the beginning of the semester, teachers would have yet to “gradually release” the students from the supports of CM, meaning that the use of CM strategies would be more apparent to the observers than they may be later in the year.

The results of the teacher observations were summarized and descriptive statistics generated prior to the teacher interviews. Following the model of explanatory sequential design, the teacher interviews included peer examination of the quantitative data. Teachers were provided with summary data describing the implementation by component and in aggregate, the implementation organized by predictor variable, and a comparison of the observed scores with the self-reported scores from the teacher reflections.

Data Collection Instruments

Observations and reflections. The *Refining Our Practices Rubric* (reproduced in the appendix) developed by E. L. Achieve was used to facilitate the collection of observational data on the use of CM critical components. The rubric has four indicators for each of the five critical components. Each indicator is evaluated on a four-point scale, with point descriptors for each indicator. According to the CM program manual (2008), the rubric has been designed and modified by E. L. Achieve and used in multiple sites across the country. The feedback from users, including teachers and coaches, has been collected to make modifications to the rubric over time. For example, a past version of the rubric included descriptors for individual items that did not explicitly build on each other: it was possible to receive a rating of 4 without first meeting the requirements of level 3. As a result, the current version of the rubric includes descriptors for level 4 items that include the phrase "in addition to level three criteria" (plus added level four criteria). The rubric items have also been reorganized to complement the rewording of the components. For example, in an upcoming version of the rubric, the items that reference writing in component (4), *Interactive Reading and Note-Taking,* are moved to the *Language for Writing Comprehension* section.

According to representatives from E. L. Achieve, although field-testing has occurred, the results have not been presented for publication in peer-reviewed journals, outlined in the program manual, or published in any sort of technical manual. As described above, the publishers opted to refine the rubric over time based on feedback from users rather than reporting on the reliability and validity of the tool. The lack of reliability and validity data for the rubric is a concern, as virtually all of the quantitative data for the evaluation of CM was derived from the rubric. Without psychometric data available, the data collected is simply assumed to be reliable and valid, which can lead to misinterpretations. Therefore, a reliability analysis was conducted using the observational data collected in this study. Limitations associated with the use of an untested instrument will be thoroughly considered in the discussion.

Teacher surveys. A survey was developed to be completed by all teachers in this study. The intent of the survey was to gain insight into the use of CM instructional practices and the reasons behind varying levels of implementation fidelity. The survey items were developed based on the work of the Learning Forward organization (formerly the National Council of Staff Development). Learning Forward developed their current standards for professional development to outline the characteristics of professional learning that lead to effective teaching practices, supportive leadership, and improved student results (Learning Forward, 2014). The *Standards Assessment Inventory* (SAI) was developed to assess the quality of professional development in schools, based on standards defined by Learning Forward (Vaden-Kiernan, Jones, & McCann, 2009). The SAI has been used in case studies that documented the use of Learning Forward standards in professional development planning and evaluation (Slabine, 2011). For example, between 2008 and 2010, 285 schools in Arkansas used the SAI to evaluate their implementation of the Arkansas Comprehensive School Improvement Plans. Their results indicated that by aligning to Learning Forward's standards, school leaders were able to understand "what areas were having an impact and what areas needed improvement." From the case study, the Arkansas Department of Education identified the evaluation of professional development as an official point of emphasis for local school leaders (Slabine, 2011).

The SAI itself is too broad and cost-prohibitive to utilize directly in this study. Instead, individual survey questions for this study were aligned with the Learning Forward standards of leadership, resources, and implementation (Learning Forward, 2014). The survey contained two sets of questions. The first set was targeted at the overall implementation process, with four questions primarily addressing leadership and five questions primarily addressing resources. The second set was four basic questions, all primarily targeted at implementation, repeated for each of the five critical components (a total of 20 implementation questions in the second part of the survey).

Additionally, the survey included open-ended questions that allowed teachers to provide as much detail as desired in their responses. The questions asked teachers to explain how well aligned the trainings were to their past instructional practices, the extent to which the training required teachers to modify their curricular materials, which elements of the training and follow-up made implementation easy, and which elements of the training and follow-up were difficult.

Interviews. Semi-structured interviews were conducted with teachers (*n* = 9) in order to gain a more in-depth understanding of their perspectives regarding CM implementation (Merriam, 2014). Interviews provided rich and meaningful data used to understand the different levels of fidelity. Maxwell (2013) suggests that interviews should be used to gain a description of the contextual details that are difficult to uncover by observation alone. Weiss (1994) and Maxwell agree in directing researchers to ask questions specific to the observations. The guidance of Weiss and Maxwell was followed in this study by providing teachers preliminary results based on observations in order to hear their insights as to why certain trends emerged.

The quantitative data was used to drive certain aspects of the interviews, such as asking teachers to explain trends. Specific pieces of data were selected that aligned to sub-questions (a) through (c) of RQ1 regarding the degree of implementation and which variables seemed to predict implementation success. Interviewees were presented with three relevant outcomes of the quantitative measures. First, the interviewees were asked to comment on the distribution of the fidelity index by component and overall implementation. Next, the fidelity index, disaggregated by primary subject area, was shown for comment. Finally, interviewees discussed a side-by-side comparison of the indices showing the observations with the scores that were self-reported during the reflections. The indices, as described in the methods of analysis sections below, were displayed as a percentage of points earned for each of the components on the rubric (e.g., all level 3 scores would be displayed as 75%).

Focus groups. Eight focus group sessions were conducted, with approximately five students each and lasting between 45 and 60 minutes. The questions were structured around the students' opportunities and use of the classroom techniques of CM practices. The students were presented with an age-appropriate definition of the goal, along with a few sample tasks for each of the five critical components. For example, the goal statement for component (5), *Academic Writing Support*, was presented as "Teachers are trying to help you write like professionals in each of your subjects." The sample task was the use of sentence frames to provide evidence for an argument. Students were asked questions like "How have you used sentence frames in your different classes?" Students were also given the opportunity to follow up on these answers with more open-ended questions, such as “Did you find sentence frames helpful in completing your assignments?” Similarly, when investigating component (3), *Oral Language Practice,* students were asked about their opportunities to talk to each other during class. For example, they were asked if they were able to choose their own groups, if they had used the “appointment clock,” and whether they were taught different techniques in (active) listening. Similar prompts and examples were provided for each of the five critical components.

Procedures

The procedures described below took place during a six-week interval beginning in early February 2015. The observations, reflections, and surveys were completed in the initial weeks, followed by the teacher interviews and student focus groups.

Observations. Each teacher was observed by one of the two district instructional coaches for one 20-minute interval of a lesson. As described above, E. L. Achieve certified the observers through additional training to support implementation. The observers were also classified as certified teachers and not as administrators, ensuring that their presence would not be used for job performance evaluations. The observers maintained confidentiality by using codes instead of teachers’ names in all documentation. The observations were preannounced but not necessarily scheduled. Teachers were able to choose from certain blocks of dates and times, but did not know the exact time of the observation. The goal of this scheduling system was to eliminate activities that would prevent observation of instructional practices, such as tests or guest speakers, while also attempting to see “regular” practice. The observers also attempted to ensure that an approximately equal number of observations occurred during the beginning, middle, and end of the class period. However, due to logistical limitations, approximately 20% of all observations occurred during the beginning third of the class, 40% in the middle, and 20% during the closing third.

Teacher reflections (self-evaluations). The teachers were asked to self-assess their typical practices by using the *Refining Our Practices Rubric*, which was distributed to every teacher via Google Forms. Each teacher was assigned a code number, allowing their evaluation data to be linked with their classroom observation without requiring individual names to be used. All information was kept confidential in order to increase confidence in the formative rather than potentially evaluative nature of this study. Teachers are more likely to participate authentically if they have confidence that their results will not be connected to their names without prior permission (Fink, 2013). The instructional coaches were the only people with a master list of names and numbers, and did not disseminate any identifiable information, as required by the University's Internal Review Board.

Surveys. The teacher surveys were distributed with the self-evaluation form discussed above. The principal of the school allocated an hour of staff development time to complete the teacher reflection and survey; however, participation was kept optional. By allotting time for survey completion, participants may have been more likely to complete the survey than if they were asked to complete it on their own time. The intent was that, through careful communication, teachers would recognize this survey as an opportunity to have their voice drive PD planning and they would take the time to respond thoughtfully. Initial participation was 30%, so a reminder email was sent asking teachers to complete the survey within a week of the allotted completion time. Following the month that the survey was open, a final open-ended survey question was sent to all teachers asking them if they would explain why they did not complete the survey, if applicable. The purpose of the open-ended question was to gain an understanding of why individuals chose to not participate in order to better understand a possible source of sampling bias (Fink, 2013). Only two teachers replied, and they simply stated that they lacked time to complete the survey. The process described above resulted in a total 43% completion rate for the survey.

Interviews. The instructional coaches provided a list of nine teachers, three from each third of the implementation distribution that was derived from the observations. The self-evaluations and surveys were not used to generate the interview list because of the low response rate. The distribution was constructed by ranking individual teachers in order by their overall implementation index, and then dividing the list into three groups. The resulting list included three groups of 10 teachers each. A random number generator was then used to select three teachers from each group for interviews. The intent was to obtain a range of responses; however, the identities of the teachers remained confidential. One hour was allocated for each interview. Interviews were conducted in person at the school.

The interviews began with a simple introduction of the purpose and overall design of the study. Teachers were reminded that the survey information would remain confidential and not be used for any job performance evaluations. The interviews followed a semi-structured script, including open-ended questions and a discussion of the quantitative data results as described above. The audio of the interviews was recorded using Audacity software and transcribed by the Casting Words Transcription Service. The transcripts were then loaded into ATLAS.ti software for analysis.

Student focus groups. Fifteen students from each grade level (6–8) were selected at random to be invited to participate in focus groups. The grade-level groups were then organized into three groups each, consisting of approximately 5 students. Students were first notified in their homeroom classes, given a written description of the study, as well as a consent form for their parents to sign. Phone and email communication was encouraged between the students' families and the principal investigator. All parents of the 45 selected students agreed to the participation, and focus groups were scheduled during homeroom to limit the disruption to instruction. The students were called to the office by the school secretary just before their scheduled focus groups. Each group met in a central conference room and sat around a circular table. The researcher facilitated the focus groups in a casual and welcoming tone, first asking each question to the group and then ensuring that each student had the explicit opportunity to respond to each question. The audio of the focus groups was digitally recorded, transcribed, and coded in the same manner as the teacher interviews.

Methods of Analysis

Research Question 1. *How successfully has the faculty of a local middle school implemented the critical components of Constructing Meaning training*? The analysis of RQ1 was based on data gathered from the *Refining Our Practices Rubric*. The rubric provided data from two data collection techniques, the observations by coaches and the reflections by teachers. These tools provided two sets of scores for each of the critical components of CM training. Each of the five critical components was evaluated using four indicators, and each indicator was measured using a four-point Likert scale (1–4). For each critical component, then, there were 16 possible points (four for each of the four indicators). The sum of the component scores yielded the overall index of fidelity. The scores from the observations and the scores for the reflections were calculated using identical techniques, and kept separate for comparison. In Chapter IV, the results are reported for both the component scores and the overall indices using descriptive and inferential statistics. The purpose of the quantitative analysis was to identify the degree to which the components had been implemented, the variation in implementation across the site, and to compare the implementation to the predetermined standard of 75%. In addition, the data gathered were used during interviews to prompt teachers to describe the nuances of the implementation, judge the overall implementation (RQ1), and reflect on the conditions that support or hinder implementation (RQ2).

Descriptive statistics were used to compare the scores assigned by the instructional coaches with the scores given by the teachers. Comparing the scores on identical sections of the rubric made it possible to determine whether the perceptions of teachers differed from the practices observed by coaches.

Research Question 2. *What are the conditions that favor or hinder a high degree of implementation fidelity of Constructing Meaning practices?* Surveys and teacher interview data were collected to answer this question. As suggested by Creswell (2014), qualitative data were analyzed as follows. Open-ended survey question responses, interview transcripts, and student focus group transcripts were organized into ATLAS.ti, a qualitative coding program. The survey and interview data were organized by teacher characteristics of subject area taught, years of experience, and overall level of fidelity as indicated by the observations. The data were coded, using both preset and emerging codes. The preset codes were based on the professional development standards identified by Learning Forward and included leadership, resources, content knowledge, student ability, and quality of training. In order to account for bias in selecting these codes, themes were added that emerged from the data analysis. The codes were further organized into a small number of themes (3–4). Themes were then interpreted and verified using the member checking technique described below.

Threats to Reliability and Validity

The methods described above introduce several threats to the reliability and validity of the data obtained. The lack of psychometric data available for the *Refining Our Practices Rubric* and the absence of inter-rater agreement and training prior to this study presents a significant threat to the quantitative data. The qualitative instruments were developed solely for the case study presented here and introduce threats to validity and reliability. In the following sections, the threats are further described and attempts to limit their impacts are identified.

Reliability. There are two significant areas of concern regarding reliability in this study. The first reliability concern stems from two different observers completing the observations. The concern is that the raters may not have consistently applied the rubric to their observations. In an attempt to increase inter-rater reliability, the observers practiced using the rubric with prerecorded video examples of lessons employing CM strategies obtained from E. L. Achieve. The observers then performed five live pilot observations together, with each classroom visit lasting approximately 20 minutes. The teachers being observed knew that the coaches would be visiting their classes during a specific week, but did not know the exact period of the day. The coaches met prior to the observations to review the rubric, watch example video clips provided by E. L. Achieve, and discuss possible “look-fors” for each indicator.

The observers followed a routine for completing and rating the teachers during the pilot observations. They would both observe the teacher, take any necessary notes, and determine ratings independently. Following the observation and independent analysis, the coaches would discuss what they observed and compare their scores. The coaches would then agree on a final rating for the teacher to be used in the reliability analysis. Results were analyzed for inter-rater reliability using percent agreement and Cohen’s kappa (Morgan et al., 2013).

The second significant reliability threat follows from the instrument. The rubric is designed so that each of the four items per component are weighted equally to provide an overall component score. The reliability threat stems from the possibility that the items do not measure the same construct. As previously stated, the developers of the rubric do not publish psychometric data for the tool. Therefore the reliability of the instrument was analyzed using Cronbach’s alpha (Cronbach, 1951; Tavakol & Dennick, 2011). The results of the reliability analysis are presented in Chapter IV.

Construct validity. Mowbray (2003), Bond (2000), and Century (2010) each present methodologies for designing and validating fidelity studies. Both Mowbray and Bond suggest using experts to determine the components to be measured. E. L. Achieve developed the components and evaluation rubric used in this study. The question then becomes, does the rubric provide a valid assessment of each implementation component? As stated earlier, the rubric used in this study has not been formally evaluated. The only indication of validity is the face validity claimed by the developers and supported by the instructional coaches involved in this study. *Face validity* refers to whether or not a measure seems or appears to be valid as determined by the individuals using it (Babbie, 2007). However, as the name implies, face validity is only a superficial indicator of validity and is not robust enough to provide significant confidence in the measures (Thorndike & Throndike-Christ, 2011). The lack of confidence in the validity of the rubric is a study limitation that will be addressed in the discussion.

Qualitative validity. Qualitative validity is related to the accuracy, integrity, and credibility of the study (Cresswell, 2013; Maxwell, 2005). The accuracy of the current study has been strengthened through triangulation of data, the integrity through negative information, and the credibility through member checking.

Triangulation. Data for the qualitative portion of the study was obtained from the following sources: open-ended survey questions, interviews with teachers, and student focus groups. Each of the data sources was targeted to explore the results of the quantitative portion of this study. The variety of data sources strengthens the validity of the study by reducing chance associations and biases of a single measure (Maxwell, 2013). For example, open-ended survey responses can indicate emerging themes that are relevant to the study. The themes can be confirmed (or refuted) through interviews and focus groups.

Discrepant information and negative cases. Qualitative studies often uncover both positive and negative information. Investigating and specifically reporting findings that may not align with researchers' desired outcomes (discrepant information and negative cases) is a key strategy in strengthening qualitative validity (Maxwell, 2013).

The inclusion of such data indicates that the researcher has presented a comprehensive analysis of the data. Generally speaking, based on personal communication with teachers, coaches, and school and district administrators, there was a desire to see a high level of implementation across all components of CM training. However, as presented and discussed in Chapters IV and V, data that suggested areas for improvements were also included in this study.

Member checking. A substantial amount of evidence derived from the qualitative surveys and interviews with teachers needed to be interpreted for analysis. The strategy of member checking was used in an attempt to prevent misinterpretation of teachers' statements. Using member checking allows participants to comment on the findings and report whether they agree with the theories that have developed (Creswell & Plano Clark, 2011). Preliminary summaries of each interview, along with developing theories, were shared with teachers. Teachers were asked to clarify and comment on the findings. Maxwell (2013) considers member checking to be the most important method for eliminating misconceptions and uncovering bias in qualitative analysis.

The research design, data collection methods, and analytic procedures described above were designed to investigate the research questions presented in Chapter II. The quantitative measures utilized in phase one provided the data to determine the degree of implementation and the variation in implementation across the site. In addition, the initial analysis of the quantitative data was used during the qualitative interviews. Finally, qualitative data were used to describe the conditions that favored or hindered implementation. The results of the study are presented in chapter IV.

CHAPTER IV

RESULTS

Chapter IV is organized by the two research questions addressed in this study. RQ1 investigated the success of implementation, through the evaluation of the implementation of critical components, the variability of implementation, and in comparison with a developed standard. RQ1 was addressed using primarily quantitative data. RQ2 investigated the conditions that favored or hindered successful implementation and was addressed using primarily qualitative data.

RQ1: Success of Implementation of CM Practices

RQ1 investigated the success of implementation by constructing fidelity indices from the observation and reflection data conducted with the *Refining Our Practices Rubric*. The investigation included a reliability analysis of the rubric and the computation of fidelity indices for the components and the aggregate.

Results of the calibration observations. Informal phone interviews with the observers revealed that they felt they were “generally on the same page” in regard to the ratings. They stated that they felt discrepancies likely occurred due to observing different parts of the class, such as one rater watching the teacher and the other focusing on student actions, rather than a different interpretation of the same observation. For example, one rater mentioned that in the first observation, she found herself focusing on a particular group of students and missed the teacher providing direct instruction related to interactive note-taking. During the debriefing, the second rater brought the missing evidence to her attention, and she agreed a higher rating would have been more appropriate. Another example was described by rater two, who began each observation by scanning the room for evidence of anchor charts displaying sentence frames, word walls, or other student aides. As a result, she did not record any of the verbal instructions or student responses that occurred during those first few minutes. The routine was therefore modified to have observers scan the room at a time when there was a lower chance of missing verbal evidence, such as during silent reading.

Quantitative analysis of the calibration data. Inter-rater reliability was estimated using Cohen's kappa. Each rater had made 100 judgments during the calibration process that resulted in a kappa of .607, indicating a moderate level of agreement (Cohen, 1960; Mchugh, 2012).

Internal Consistency

Internal consistency was estimated using Cronbach's alpha (α), inter-item correlations, and recalculation of Cronbach's alpha with each item removed. Following the complete analysis of the reliability data, one item was excluded from all analyses.

Cronbach's alpha for each component. The *Refining Our Practices Rubric* includes sets of items for each component separately. In reality, the rubric is actually a collection of five rubrics, one for each critical component. Therefore, the rubric was treated as five unique tests, one for each component, with four items each (n = 4) when calculating Cronbach's alpha.

Alpha values ranged from .290 to .837. A general rule of thumb (Gliem & Gliem, 2003) suggests that an alpha value of > 0.9 is excellent, > 0.8 is good, > 0.7 is acceptable, > 0.6 is questionable, > 0.5 is poor, and < 0.5 is unacceptable. As can be seen in table 1, scores on component (1), *Understanding Backward Design* (UBD), and component (2), *Language as a Part of Content Teaching* (LPCT), had alpha values corresponding to "good" reliability (α = .803 and α = .837, respectively), while component (3), *Oral Language Practice* (OLP), and component (4), *Interactive Reading and Note-Taking* (IRNT), were on the border of "questionable/poor," with α = .606 and α = .597, respectively. Component (5), *Academic Writing Support* (AWS), demonstrated the lowest reliability at α = .209.

Inter-item correlation analysis. An inter-item correlation analysis for each item within each component was also conducted. It was expected that all items for a particular component would show an acceptable agreement with all other items. The highest possible inter-item correlation may not be the most desirable situation. Correlations that are too high may indicate repetition between items and a narrow illustration of the desired construct (Tavakol & Dennick, 2011). Various but similar "rules of thumb" appear in the literature. Generally, inter-item correlations are acceptable above 0.25 and below 0.70 (Briggs & Cheek, 1986; Clark & Watson, 1995). As can be seen in table 1, all items for component (1), *Understanding Backward Design,* and component (2), *Language as a Part of Content Teaching,* showed positive and acceptable correlations between items. The items measuring component (3), *Oral Language Practice,* all displayed positive correlations, although items three and four correlated with each other at a low level (.123). Likewise, *Oral Language Practice* items three and four correlated below the .250 threshold, at .215 and .222, respectively. Component (4), *Interactive Reading and Note-Taking,* items one and four showed a negative correlation. *Interactive Reading and Note-Taking* item four's mean correlation between items also fell below the threshold at .175. The indicators of component (5), *Academic Writing Support,* showed the lowest correlations between items. Each *Academic Writing Support* indicator showed at least one negative relationship with the other items. *Academic Writing Support* item three performed particularly poorly, having a negative mean overall correlation between items (-.021).

Cronbach's alpha with items deleted. Following the correlation analysis, Cronbach's alpha was recalculated for each component while removing each item and including only the three remaining items. The purpose of recalculating was to determine whether the overall alpha for the component was increased as a result of removing one of the four items. An increase in alpha values suggests that removing the item would be beneficial for component reliability. Table 2 displays all alphas with items deleted compared to the original alpha that included all items.

Removing a particular item led to a decrease in the alphas in 13 of the 16 possible cases within components (1) through (4), OLP, UBD, LPCT, and IRNT, suggesting that those 13 items remain in the analysis. The three alphas that did rise as a result of the removal did so minimally; removing LCPT item four caused an increase of 0.23; OLP item three, an increase of 0.09; and IRNT item four, an increase of 0.28. Therefore, the three items remained in the analysis. In component (5), AWS, removing either item two or three would raise the alpha for the component, to alphas of 0.32 and .62, respectively.

Revisions due to reliability analysis. Because AWS item three substantially lowered the level of reliability, AWS item three was removed from all further analysis. As can be seen in table 2, the change resulted in a revised Cronbach's alpha value of .619, compared to the original value of .290.

Table 1

Summary of inter-item correlations within each critical component

	Minimum	Maximum	Mean
UBD Item One	.331	.505	.442
UBD Item Two	.402	.685	.526
UBD Item Three	.505	.685	.605
UBD Item Four	.331	.625	.452
LPCT Item One	.458	.864	.633
LPCT Item Two	.375	.864	.600
LPCT Item Three	.520	.578	.553
LPCT Item Four	.375	.458	.451
OLP Item One	.212	.512	.320
OLP Item Two	.286	.512	.377
OLP Item Three	.123	.286	.215
OLP Item Four	.123	.333	.222
IRNT Item One	-.019	.615	.273
IRNT Item Two	.223	.431	.305
IRNT Item Three	.112	.615	.329
IRNT Item Four	-.019	.431	.175
AWS Item One	-.058	.831	.340
AWS Item Two	-.025	.247	.068
AWS Item Three	-.058	.012	-.021
AWS Item Four	-.025	.831	.270

Table 2

Revised Cronbach's alpha values with specific items deleted

Understanding Backward Design (Original Cronbach's alpha = .803)	
	Alpha with item deleted
UBD Item One	.799
UBD Item Two	.736
UBD Item Three	.672
UBD Item Four	.787
Language as a Part of Content Teaching (Original Cronbach's alpha = .837)	
LPCT Item One	.721
LPCT Item Two	.747
LPCT Item Three	.804
LPCT Item Four	.860
Oral Language Practice (Original Cronbach's alpha = .606)	
OLP Item One	.499
OLP Item Two	.399
OLP Item Three	.614
OLP Item Four	.599
Interactive Reading and Note-Taking (Original Cronbach's alpha = .597)	
IRNT Item One	.524
IRNT Item Two	.504
IRNT Item Three	.420
IRNT Item Four	.625
Academic Writing Support (Original Cronbach's alpha = .290)	
AWS Item One	-.100
AWS Item Two	.318
AWS Item Three	.619
AWS Item Four	.019

Fidelity of Implementation Index by Component

As described in Chapter II, determining the degree to which implementation varied across the components would be used to evaluate the implementation of CM practices. The index of fidelity was constructed for each component by calculating the percentage of points earned from the four items. All of the components, except *Academic*

Writing Support (AWS), have 16 points possible (four points from each of the four items). AWS had 12 total points possible, since item three was removed from analysis following the reliability analysis.

Table 3 displays the percentage of implementation listed by component. Percentages were used instead of raw scores due to the varying total possible score across components. Overall, *Understanding Backward Design* was implemented with the highest level of fidelity (66.67%). In contrast, *Oral Language Practice* was implemented at the lowest level (40.83%).

The level of implementation also varied within the components. As can be seen in table 3, *Oral Language Practice* was implemented with the least variance (SD = 13.10) between teachers. In contrast, *Language as a Part of Content Teaching* (SD = 21.17) was implemented with the most variance between teachers.

Overall index of fidelity. The overall index of fidelity is simply the percentage of all points awarded on the rubric excluding AWS item three. The minimum fidelity index, as a percent, was 34.21%, with the maximum being 78.55%. The mean and median scores were 51.40% and 48.69%, respectively (SD = 11.51).

Index of Fidelity by Predictor Variable

Fidelity rates were examined as a function of the predictor variables described in Chapter III: years of teaching experience, primary subject area, and time latency since training.

Years of teaching experience. The participants included teachers ranging from 1 to 25 years of experience, with a mean experience of 9.85 years. The indices were

Table 3

Index of fidelity by critical component (n = 30)

	Minimum (%)	Maximum (%)	Mean (%)	SD
Understanding Backward Design	37.50	100.00	66.67	17.70
Language as a Part of Content Teaching	31.25	100.00	55.00	21.17
Oral Language Practice	25.00	68.75	40.83	13.10
Interactive Reading and Note-Taking	25.00	75.00	50.00	13.23
Academic Writing Support	25.00	83.33	42.22	18.17

analyzed for a relationship to years of experience. Table 4 displays positive correlations between years of teaching experience and the overall index as well as four of the components, although only one was statistically significant at the .05 level (IRNT, *r* = .455, *p* = .012). The only component to show a negative correlation was *Oral Language Practice* (*r* = -.068). These results suggest that for all of the components except *Oral Language Practice*, as years of teaching experience increased, the level of implementation also increased.

Subject area taught. Fidelity indices were compared between teachers' primary subject area. Table 5 displays the number of teachers in each subject area. The small group sizes are further addressed in Chapter V. As can be seen in figure 3, humanities teachers implemented AWS (62.5%) and IRNT (54.7%) more than the other subject

Table 4

Pearson's correlation between year of teaching experience and fidelity of implementation

	UBD	AWS	IRNT	LCPT	OLP	Overall
r	.320	.318	.455*	.161	-.068	.339
Sig. (2-tailed)	.084	.087	.012	.396	.719	.067

Note: * = significant at the .05 level.

areas. Special education teachers demonstrated the highest implementation in OLP (52.1%) and LPCT (68.8). In UBD, physical education (PE) was the highest (81.25%), although there was only one teacher in the group. Conversely, art teachers showed the lowest level of implementation in all components, except IRNT, in which the PE teacher showed the lowest level of implementation.

Table 5 displays the overall index of fidelity by primary subject area. As can be seen, the humanities and special education teachers showed the highest degree of implementation at close to 60% each, while science and art showed the lowest at 44.57% and 39.47%, respectively. The statistical significance of the differences was tested using a one-way analysis of variance.

Analysis of variance by subject area taught. The descriptive statistics revealed variability in implementation by teachers in different subject areas. A one-way ANOVA was conducted to compare teachers' primary subject area with the implementation index means for each of the critical components, as well as the overall index, in order to determine whether there were statistically significant group differences. Due to the small sample size, teachers were combined into three groups for the ANOVA: group one was humanities teachers (*n* = 6), group two was science or math (*n* = 18), and group three was

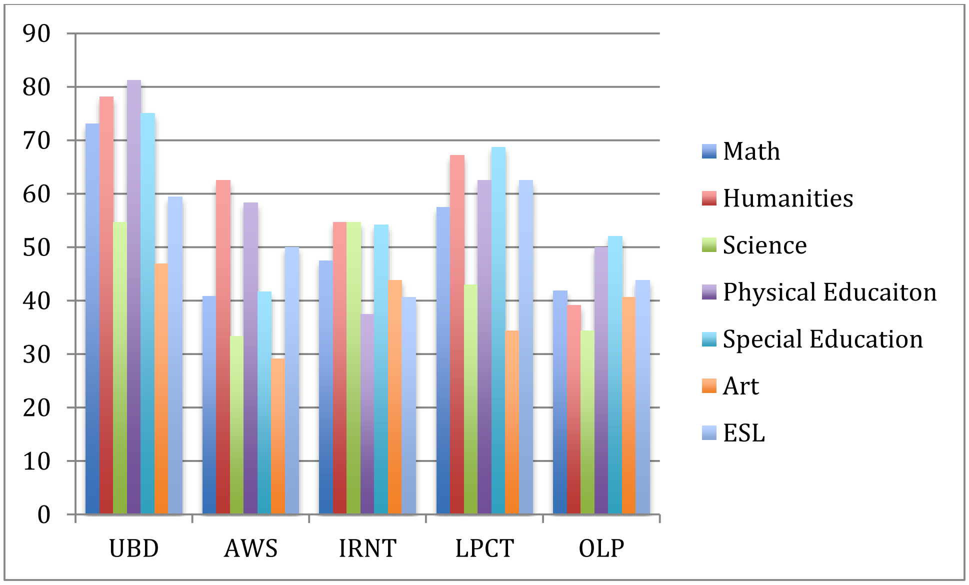


Figure 3. Implementation of each critical component by subject area.

Table 5

Index of fidelity of the overall intervention by teachers' primary subject area

	n	Index (%)	SD
Math	10	52.76	11.33
Humanities	4	60.20	13.13
Science	8	44.57	3.62
Physical Education	1	57.90	**
Special Education	3	59.21	19.74
Art	2	39.47	7.44
ESL	2	51.32	5.58

electives (*n* = 6). Only one statistically significant result was obtained. There was a significant relationship between subject area and the implementation of *Academic Writing Support* [F(2, 27) = 3.523, MSR = 28.873, *p* = 0.044].

Table 6 shows that the post hoc comparisons using the Tukey HSD test indicated that the mean score for ELA/humanities teachers was significantly different than for science/math teachers. There was no statistically significant difference between ELA/humanities and elective teachers or between science/math and elective teachers.

Table 6

Tukey HSD post hoc test

		Mean Difference	Std. Error	Significance
ELA/humanities	Math/science	20.83333*	7.90509	.036
	Electives	18.05556	9.68172	.168
Electives	Math/science	2.77778	7.90509	.934

Note: * = significant at the 0.05 level.

Latency since training. The participants each completed the training within a two-year window. In order to investigate differences in implementation due to the amount of time elapsed since the training, the teachers were organized into three cohorts. Cohort one completed the training during the 2013–2014 (*n* = 9) school year, cohort two completed it during the following summer (*n* = 8), and cohort three (*n* = 13) during the 2014–2015 school year.

Fidelity indices were examined by critical component and overall implementation. As can be seen in Table 7, each cohort showed a greater level of fidelity *Understanding Backward Design*. In the overall index, however, there was a minimal (< 4%) difference in the implementation index among cohorts. Additionally, a one-way analysis of variance did not show a statistically significant relationship between the different training cohorts and the implementation of any specific component or overall.

Table 7

Index of fidelity organized by training cohort

	n	UBD		AWS		IRNT		LPCT		OLP		Overall	
		(%)	SD	(%)	SD	(%)	SD	(%)	SD	(%)	SD	(%)	SD
Cohort 1	9	74.31	17.52	38.89	22.05	47.22	15.02	53.47	24.03	40.28	11.31	51.46	13.23
Cohort 2	8	67.19	22.60	47.92	16.52	43.75	12.05	65.63	16.02	41.41	17.01	53.45	10.93
Cohort 3	13	61.06	13.30	41.03	16.83	55.77	10.96	49.52	20.96	40.87	12.66	50.10	11.35

Teacher Reflection and Survey

The electronic document that contained both the teacher reflection and the survey was distributed to all participants. Thirteen teachers (43%) responded during the month that the tool was open. Five of the respondents had been teaching 16 or more years, four between 11 and 15 years, and four between 6 and 10 years. Five primarily teach math, five language arts, and three science. Eight teachers participated in CM training during the current school year, four the prior year, and one took the training two years prior. One teacher maintains a state teaching license in English for Speakers of Another Language (ESOL).

Teachers' self-scoring on the *Refining Our Practices Rubric*, by component. The fidelity indices displayed previously in this chapter were calculated from the observations conducted by the instructional coaches. The observers could only score the rubric based on what they directly saw or heard, and could have missed evidence of implementation that may have occurred prior to or following their visits. The teachers' self-reflection of the rubric was analyzed in order to include evidence of implementation that may not have been otherwise observable. An index of fidelity was calculated from only reflection scores using the same technique as for the observational data, including dropping AWS item three.

Table 8 displays the indices by component as well as overall. As outlined in the table, teachers scored themselves the highest in UBD and LPCT, 63.94% and 65.86%, respectively. Although the numerical values for the indices are different, those were also the components scored highest during the observations. A comparison of the observed and self-reported scores, along with a test of statistical significance, is presented below.

Although the sample size (*n* = 13) was too small to achieve sufficient statistical power, the observations and participant reflections were analyzed to determine whether any trends emerged that could be recommended for a future, more comprehensive study. A paired sample *t*-test was used to compare the means for each critical component and the overall index. As illustrated in figure 4 and detailed in table 9, the means between observed scores and reflection scores were statistically different ($p < .05$) for AWS (MD = -14.74, $p = .04$), IRNT (MD = -14.01, $p = .03$), and OLP (MD = -13.08, $p = .01$).

RQ2: Conditions That Favor or Hinder Successful Implementation

The goal of RQ2 was to determine the conditions that supported or hindered the implementation of CM practices. The investigation of RQ2 included exploring the teachers’ perceptions of the overall training and their perceptions of implementing the practices at the classroom level, through both open- and closed-ended survey questions. The surveys yielded traditional qualitative data as well as numerical summaries of responses. Additionally, teachers’ perceptions of the initial analysis of quantitative data and CM concepts in general were discussed during semi-structured interviews. Finally, focus groups with students were conducted to understand their experiences with CM practices.

Table 8

Index of fidelity based on teacher reflections

	Minimum	Maximum	Mean	Median	SD
Understanding Backward Design	50.00	75.00	63.94	62.50	9.59
Language as a Part of Content Teaching	43.75	81.25	65.86	62.50	15.43
Oral Language Practice	43.75	81.25	58.65	56.25	18.31
Interactive Reading and Note-Taking	56.25	93.75	10.92	62.50	9.70
Academic Writing Support	33.33	91.66	53.20	50.00	15.79
Overall Implementation	47.36	88.15	62.45	61.84	11.49

Teachers' Perceptions of CM Training

Nine of the survey questions were closed-ended, asking teachers to evaluate their overall experience with CM trainings. Thirteen (42%) of the teachers participating in the study chose to complete the survey and reflection. As shown in table 10, 100% of respondents felt that the use of CM practices would have a positive impact on student outcomes. Seventy-seven percent felt that the leadership at the school was able to

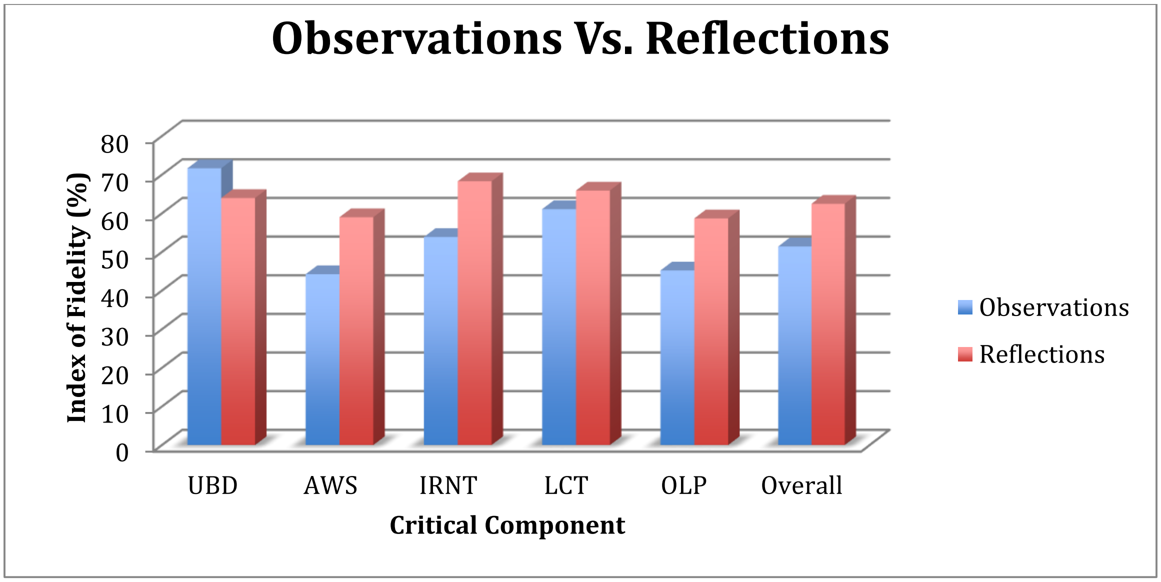


Figure 4. Comparison of scores from observations and reflections.

Table 9

Paired sample t-test of the observation and reflection indices

	Paired Differences							
	Mean Difference	SD	95% Confidence Interval of the difference		Correlation	*t*	df	Sig. (2-tailed)
			Lower	Upper				
UBD$_{obs}$ - UBD$_{ref}$	7.69	18.07	-3.22	18.61	0.12	1.53	12	0.15
AWS$_{obs}$ - AWS$_{ref}$	-14.74*	23.11	-28.71	-0.78	0.06	-2.30	12	0.04
IRNT$_{obs}$ - IRNT$_{ref}$	-14.04*	15.60	-23.85	-5.00	0.18	3.33	12	0.01
LPCT$_{obs}$ - LPCT$_{ref}$	-4.08	31.16	-23.63	14.02	-0.28	-0.56	12	0.59
OLP$_{obs}$ - OLP$_{ref}$	-13.08*	19.41	25.19	-1.73	0.26	-2.50	12	0.03
Overall$_{obs}$ - Overall$_{ref}$	-8.14	16.91	18.36	2.07	-0.03	-1.74	12	0.11

Note: * = significant at the 0.05 level.

demonstrate that CM was a priority. Accordingly, 92% of respondents disagreed that the school should be setting different priorities for professional development. Ninety-two percent of respondents also felt that they were able to make connections between the collaborative “learning team” model and the practices associated with CM. However, only 76% agreed that they were able to collaborate with the learning team on CM practices. Ninety-two percent of respondents disagreed that CM practices are too time-consuming to implement; however, only 39% felt that there was time during the school day to work on implementation.

Teachers' perceptions of each critical component of CM practices. Each respondent was asked six Likert scale questions for each of the five critical components. The six questions were identical for each component. In all components, the majority of respondents (> 75%) felt that the practices associated with the critical component were aligned with the subject area they teach, relevant to the students' needs, and able to be fully implemented within the next two years. There was disagreement regarding whether full implementation would cause a significant change to teaching practices. Sixty-two percent of respondents felt that *Language as a Part of Content Teaching* would require significant change to current teaching, and only 31% responded similarly to *Interactive Reading and Note-Taking.*

Open-ended survey questions. The survey included four open-ended questions regarding the teachers' perceptions of the impact of CM training. Teachers were asked to provide evidence of connections to past practice, changes that were required as a result of the training, elements that made the practices relatively easy to implement, and elements that made the practices relatively difficult to implement. Of the thirteen teachers completing the survey, twelve answered the questions (although answers were required for submissions, one teacher simply put an "x" in each of the response spaces to move on). Example responses are shown in Table 11. Specific quotations are also provided and interpreted in Chapter V.

Table 10

Summary of closed-ended survey questions

	Strongly Agree (%)	Agree	Disagree	Strongly Disagree (%)
I was able to make direct connections between CM training and my learning team.	46	46	1	0
I was able to collaborate with other members of my learning team regarding CM practices.	46	30	23	0
My school leaders demonstrate that CM practices are a priority.	15	62	23	0
My school leaders we able to adequately allocate resources needed for CM implementation.	15	46	30	8
In my school, there is time available to me, during the school day, to plan for CM implementation.	0	39	46	15
My input was solicited on the allocation of resources (time, consultation, learning materials) for CM implementation.	0	31	62	8
I anticipate that my use of CM practices will have a positive impact on student outcomes.	31	69	0	0
I feel CM practices are too time-consuming for me to implement.	0	8	77	15
I feel the school should have different priorities for professional development than CM.	0	15	77	8

Teacher Interviews

Nine open-ended interviews were conducted with teachers participating in the study. To select the interviewees teachers were listed by their level of implementation, and three teachers from each third of the distribution were selected at random. Of the original nine selected, four of the teachers declined to participate, and alternates were selected at random.

A protocol of open-ended questions was followed for the interviews. The protocol was developed to uncover perceptions that would be useful in the analysis of the two overarching research questions. The interviews were scheduled to take place over an hour, taking place in the teacher's classroom. The protocol began by asking teachers about their perceptions of the training itself, the systems and structures in place at their school relevant to implementation, and how the training has affected their practice. The second portion of the interview asked teachers to comment on the initial analysis of the quantitative data, both from the observations and reflections.

The teacher interview data resulted in themes related to four general categories of comments: collaboration, resources other than time, time as a resource, and general perceptions. Each category of responses yielded groups of emerging themes for analysis. Table 12 displays the frequency of at least one mention of the theme per interview.

Student Focus Groups

Seven of the 45 students selected, as described in Chapter III, did not attend their assigned focus group, presumably due to school absence on the scheduled day of the group. The resulting group consisted of 13 sixth graders, 14 seventh graders, and 11

Table 11

Example responses to open-ended survey questions

Q1. In what ways were the practices of CM aligned to your practices prior to the training?

> In the more broad sense, CM does many things that many new "initiatives" claim to do but don't: it truly does take best practice in terms of instruction (along with the pedagogical framework behind instruction) and streamlines specific ways to more explicitly implement these best practices.

Q2. In what ways did you need to modify to your curricular materials to implement CM practices?

> CM gave me really good supports or structures to explicitly teach the language I was wanting students to use. I now (occasionally) add sentence frames to my lessons and worksheets especially if I am focusing on EXPLAINING, or justifying a solution. I write frames as a part of my objectives daily.

Q3. Please describe elements of the training and follow-up that made implementation relatively difficult.

> Time to collaborate and create lessons that daily incorporate the strategies of CM. Time to digest and recognize more quickly how to change my lessons and instruction to more intentionally teach using CM strategies. Time necessary in limited class periods to instruct students in the use of all strategies, finding supportive math text readings (and the time to implement in class).

Q4. Please describe elements of the training and follow-up that made implementation relatively easy.

> Access to others that have completed the training for help. Collaboration with my learning team, which chose increased student talk and writing using academic language as our goals.

eighth graders. Twenty were male and eighteen were female. Thirty-eight percent were designated as English Language Learners (either active or monitored).

Each group of students was asked if they knew about Constructing Meaning. CM was described as a training opportunity for their teachers that helped teachers think of different ways to get students to read, write, and talk to each other. Students then looked

Table 12

Frequency of reference to emerging themes in teacher interviews

	Collaboration			General Perceptions of Training			Resources other than time		Time as a resource	
Theme	With job-alike teachers	In interdisciplinary groups	With CM trainers	Perceived effect on student achievement	Alignment to other priorities (positive)	Alignment to other priorities (negative)	Materials	Follow-up opportunities	Time with students	Time for development
Rate of Occurrence (*n* = 9)	9	7	6	9	5	4	6	6	6	9

at five sets of examples of types of assignments or tasks that teachers may have asked them to do, one set for each of the CM critical components.

None of the students had heard the terms “CM” or “Constructing Meaning.” Similarly, none of the students reported hearing of a specific training or a new or different way to have students read, write, or talk to each other.

During the demonstration of examples of types of assignments for each of the critical components, the only ones students recognized in all of the focus groups were sentence frames (LPCT) and A/B partners (OLP). Students in each group were able to identify clearly when and where they used sentence frames. Humanities classes were mentioned more frequently; however, all subjects were mentioned at least once. A/B partners was also clearly used throughout the subject areas; however, students only described it as a way to organize students into groups, rather than a way to assign different tasks to different groups members. Likewise, students were not able to explain why a particular student was assigned as “A” or “B,” other than random selection.

Interestingly, three of the six focus groups referenced Cornell notes as a structure for *Interactive Reading and Note-Taking*. Cornell notes are a specific style of note-taking, requiring students to engage with their notes frequently over the course of the day and weeks following. Although Cornell notes are an example of *Interactive Reading and Note-Taking*, they are also a significant component of AVID practices, which are also used at the school. Therefore, it was impossible to determine whether students’ exposure to Cornell notes was a result of CM or AVID trainings.

CHAPTER V

DISCUSSION

In this chapter, the results of the quantitative and qualitative findings are discussed. Following the discussion of results, the conditions that favor or hinder successful implementation of CM training are outlined. The results are compared to prior related research. At the conclusion of the chapter, study limitations and implications for future research are presented.

Discussion of RQ1: Success of Implementation

The *success* of implementation is a multifaceted concept. In the case of the implementation of CM practices, success was evaluated by the level of implementation fidelity, the variation of implementation between the components, and a comparison of the level of implementation to expected thresholds.

Success of the implementation of the critical components. As described in Chapter II, an accepted approach to measuring implementation fidelity involves defining an intervention by its critical components and measuring the implementation of each. The results of the teacher observations indicate that UBD (67%), LPCT (55%), and IRNT (50%) were implemented to the greatest degree, while AWS (42%) and OLP (41%) were implemented with the lowest degree of fidelity. The teachers' reflections on the rubric revealed similar results, although with different indices, resulting in a different order. IRNT and LPCT were reported at 66% and 64%, respectively. OLP and AWS were also reported lowest by the teachers at 58% and 53.20%, respectively. Considering this data, it is reasonable to conclude that UBD, LPCT, and IRNT were implemented more successfully (with greater fidelity) than OLP and AWS.

There is a lack of comparative research in the literature regarding the implementation of the critical components of CM training. However, in published evaluations of other interventions, the variation in the implementation of specific components is to be expected (Dusenbury et al., 2003; McKenna et al., 2014; Mowbray et al., 2003). In a larger but similarly designed study, McHugo et al. (2007) used a mixed-methods design to investigate the implementation fidelity of the five critical components of the National Implementing Evidence-Based Practices Project. The researchers evaluated fidelity of implementation across 53 different sites in eight different states. Similar to the current study, researchers used observations from multiple raters as the source of their quantitative data. Implementation data was collected four times (every six months) over a two-year period. At the first six-month point following the initial implementation, the component implementation fidelity rates ranged from a low score of 20% to a maximum score of 80%. The size of the variation between components did lessen over time, to a low of 30 percentage points between the high and the low scores (55% on the low end and 85% on the high end). These qualitative findings indicated that the components requiring “simple structural” changes were implemented with greater fidelity than those requiring changes to the “expertise” of the practitioners (p. 1283).

The researchers in the McHugo study used qualitative interviews to explain their quantitative findings. Four conditions were identified as influencing the varying levels of implementation across the research sites. The researchers identified leadership, prioritization of implementation, complexity of implementation, and the role of the trainer as areas of focus for implementation. Leaders were directed to be actively involved in the implementation process and to seek out direction from the providers to identify areas where support was needed. Researchers suggested that prioritization of implementation stems from an analysis of *why* the intervention is needed. They proposed that without a clear understanding of the purpose of the intervention, providers would be less likely to implement with fidelity. The actions of the providers are one element that makes implementation a complex process, as do the actions of the recipients, and challenges associated with resource allocation. The individual philosophies of the providers, and the range of talents that they have, were found to influence implementation. Likewise, the recipients of the interventions will have varying degrees of enthusiasm for or acceptance of the intervention. Finally, the availability of resources is likely less than the overall need, requiring careful allocation plans that include underfunding certain areas. The role of the trainer was also identified as critical to implementation. The trainer role is complex because they must have a strong knowledge of the intervention itself, while also understanding the local context enough to apply the knowledge most effectively (Torrey, Lynde, Gorman, 2005).

Similarly, the current study was designed to find out *why* certain components were implemented with greater fidelity than others. The findings of the current study were similar to the McHugo study. The implementation data by component were shared with teachers during the semi-structured interviews. The comments were analyzed, looking for explanations for the variation between components. As the interviews progressed, some themes were consistently mentioned.

One common theme was a lack of knowledge of the components themselves. Although all of the teachers were able to speak to specific strategies or assignments that had resulted from CM training, none were able to articulate the components by name or even by description. During each of the interviews, following this realization, teachers were provided with a card from CM that describes the goal of and gives examples of each of the components. Using the card as a refresher, teachers were able to speak to the various components and describe possible explanations for the variation in implementation fidelity. Although the teachers were able to discuss the components using the card as a prompt, their lack of memory of the components indicated to the researcher that they did not use the components as a way to implement CM practices.

Teachers consistently attributed the higher implementation of *Understanding Backward Design* to the emphasis that the school's district places on "standards based learning" and the required use of "learning targets." The school district has organized virtually every course taught in the district by learning targets, which are specific, student-friendly statements aligned to a particular state standard. These targets serve as the outcome measures for the class and are reposted on all forms of progress communication. As one teacher described, the process of working backward from the outcome is already built into her practice:

> I told students what they're going to do. What do you want the kids to learn by the end of the class? That's why backward design is so high. We're so target-focused. It's like this is what I need to get the kids [to do]. What is it that I need to do? How can I set up that, and how can I use Constructing Meaning to help me get where I'm going?

The higher implementation of *Language as a Part of Content Teaching* and *Interactive Reading and Note-Taking* likely resulted from the belief that these components supplement the content of the class. In *Language as a Part of Content Teaching*, teachers have students interact with the vocabulary of the course of study. Interacting with the vocabulary of the course was not reported as a new practice. As one teacher stated, “I already had some basis for academic vocabulary instructions, so CM just gave you more tools around the same conceptual understanding.” Other teachers made reference in the surveys and interviews to the quality and availability of materials. The reference to the materials indicates that teachers perceived the CM strategies with these components as *tools* to help them teach their content. The perception of the materials as helpful tools contrasts with other components that were seen as *additional* requirements not directly relevant to the content learning goals of the teachers’ classes.

There is consensus in the literature that teachers that find a relevant connection between the professional development and their teaching assignments are more likely to implement the new practices in their classroom (Darling-Hammond & Wei, 2009; Garet et al., 2009). However, less agreement exists regarding the difference between subject-specific relevance, such as presenting math methods to math teachers, and school-wide goals, such as literacy across the curriculum (Echevarria et al., 2011). The implementation of CM practices clearly followed the latter by providing specific language instruction to all teachers.

In contrast, *Academic Writing Support* was generally perceived as a method to teach writing. Although teaching writing may not be new to language arts teachers, it can present challenges to those in other disciplines. Teachers from other content areas commented that they were “not writing teachers,” and that the role of writing instruction was “one more thing that [they] need to cram in.” The perception by certain teachers that writing instruction is not a core part of their job likely contributed to the lower implementation of *Academic Writing Support.*

The challenge of integrating writing into non–language arts classes is not new (O'Brien et al., 1995; Vacca & Vacca, 1989). O'Brien and colleagues (1995) presented findings that indicated reluctance to integration stems from teachers failing to see the benefits to their primary instructional objectives. Teachers are less likely to implement changes that they do not perceive as having direct impact on their classroom objectives. Recently, however, the Common Core State Standards (CCSS) have identified content area literacy as one of the "key instructional shifts" (Bennett & Hart, 2015). The premise of the shift is the inclusion of subject-specific literacy standards throughout the curriculum. If the "shift" does occur with teachers in all subject areas, then writing strategies will be, by default, relevant to every subject area. The defined relevance would increase the likelihood that writing strategies, such as those in *Academic Writing Support*, would be implemented by teachers across the different subject areas. However, teachers' perceptions of the CCSS shift vary widely and it is not yet known whether teachers will become more accepting of "literacy across the curriculum."

In August of 2014, Gallop conducted a poll of 854 randomly selected teachers from 43 states and the District of Columbia (Saad, 2014). The teachers were asked to respond to questions asking their perception of the CCSS. The overall perception of teachers was split, with 41% responding positively and 44% negatively. The poll did find variation in perception based on level of implementation. Teachers who reported that they worked in schools that had implemented all of the standards were more likely to indicate positive perceptions (61%). A similar poll conducted by the Bill and Melinda Gates Foundation (2014) asked teachers to respond to questions regarding their overall enthusiasm for the standards and their perceptions of the impact the standards were having on their students. Of the 1,676 pre-K to 12th grade teachers that responded, 84% teachers with at least one year of implementation reported being enthusiastic about the CCSS. Similarly, 53% reported seeing positive impacts on students attributable to CCSS implementation.

The positive trend between level of CCSS implementation and teachers' positive perception gives reason to believe that the "literacy across the curriculum" shift will be seen in teachers' practices. In the current study, teachers reported that writing is still thought of as an extra or supplementary piece of the curriculum and not a direct learning target. If teachers at the school in this study do adopt the shift similar to the teachers responding to the Gallup Poll, they may be more likely to implement literacy strategies, such as those presented in CM training, in all subject areas.

The variation in implementation was also found in *Oral Language Practice*. This component was specifically described by 7 of the 12 teachers who responded to the survey as a shift in practice because they are essentially asking students to talk more, which is in contrast to the idea of students sitting quietly and waiting to speak until called upon. As described by one teacher, "A lot of times, teachers are trying to get the kids to be quiet, and this strategy would be asking them to talk more." Another teacher echoed this sentiment:

> For me, I'm trying to get them to be quiet. If I open the floodgates of letting them talk again, oh boy. I wouldn't get it back. This is a really chatty group. Also, it's contrary to what it feels like we want them to be quiet to impart whatever we're doing. If I hear noise, it's very hard to distinguish what the noise is. It is on-task or off-task?

Another teacher also described the challenge of determining whether student talk was on- or off-task behavior. It became apparent that teachers felt more comfortable teaching, and holding students accountable, in silent working conditions. The teachers did not seem to feel as comfortable designing lessons that taught the students how to talk. As one teacher said:

> It is still really hard to get students to get out of their colloquial talk and use academic language, and it's also hard as teachers to model that. I for sure think oral language is the hardest one to teach. . . . That has probably been the thing that's been the hardest. I think it didn't easily fit into the way I teach, so I have to sit down and say, "How am I going to work this in?" because I want to.

The lower implementation of *Oral Language Practice* illustrates the challenge of increasing student talk that has been described in the literature. Mitchell (2008) described 12 distinct classroom conditions that teachers should have in place to increase the amount of student talk in the classroom. The conditions included abstract concepts, such a creating environments conducive to risk-taking and fostering independent decision-making. DeWitt & Hohenstein (2010) go on to describe increasing student talk as particularly challenging for secondary teachers. The researchers report strong relationships with students as a criterion for increasing student talk. However, teachers' daily student loads of 150 to 200 made relationship-building a difficult task. The teachers in the current study experienced similar students loads, with a mean of 165 students (SD = 8.3). The takeaway is that teachers understand the elements of *Oral Language Practice* and see its value for students. However, they are either unable or unwilling to create the classroom environments where student talk is abundant.

Contextual relevance of the components. The varying implementation of the components seems to have resulted from the contextual relevance the specific practices had with individual teachers. The closer the professional development components directly apply to the specific practices of individual teachers, the more likely the teachers are to implement them as intended (Darling-Hammond & Wei, 2009; Opfer, 2011). The necessity of connecting the professional development learning objectives to teachers' daily practices must be considered by training providers and school leaders. If the PD will be presented school-wide, as in the case of CM, common agreements, such as a school-wide focus on literacy, ELL achievement, or student behavior should exist prior to implementation (Guskey, 2002; O'Brien et al., 1995; Sugai & Horner, 2002b). The middle school and its parent district include achievement of all students, specifically historically underrepresented students (including ELLs), as organizational goals. However, there is a lack of specificity as to the methods and techniques that should be used to achieve the goals. School leaders should be more explicit about the expectations of CM implementation.

Success of the overall implementation. Following the results of the component analysis, the overall analysis yielded similar results. Although the index from the observations (51%) and the reflections (62%) were different, they both indicate that the program has not been fully implemented. Although full (100%) implementation would be the ideal goal, the comparative target for this study was set at 75%, based on the descriptors in the rubric. Clearly, the implementation data collected indicates that the school has not yet met the minimal threshold of success defined by this study. However, the threshold of 75% implementation fidelity was determined for the formative purpose of this case study and was not presented to staff as an administrative expectation or a publicized goal. Because the target was not publicized, teachers were not able to use it as a target. In fact, teachers were not provided with any desired target of implementation fidelity.

The level of implementation was addressed during the teacher interviews. Every participant was asked to provide an informed opinion of acceptable level of implementation, within the current context of the school. Of the nine interviews, seven provided a number, while two did not feel they could make an informed response. Of the seven responses, there was a low of 25% and a high of 80%, with the average being 60%. In each of the responses, there was a tone of assumption that the teachers would continue progressing toward higher implementation fidelity.

An indirect indication of implementation fidelity arose from the teacher surveys. The survey asked whether teachers felt that the critical components could be implemented fully within the next two years. Although there was variation in how significant the changes required would be for individual teachers' practices, 100% of respondents agreed that each component could be implemented within two years. Combining the survey results with the indices of implementation could lead to a conclusion that implementation is progressing and will continue to grow. However, there is also counterevidence. The analysis of observational data did not reveal any consistent variation in implementation among the different training cohorts—meaning that latency or recency of teacher training had no observable effect. Because of this, implementation may not be on an upward trajectory, and an implementation plateau may have occurred. Continued analysis through a longitudinal study is needed to determine whether implementation will continue to rise. The recommendations for support are presented at the end of this chapter, and resources for continued growth would be required from school leaders.

The possible implementation plateau in the current study is similar to the National Implementation of Evidence Based Practices evaluation where researchers found time to be a statistically significant predictor of implementation from the baseline through 12 months, but not in the final 12 months of the study (McHugo et al., 2007). Furthermore, four of the five components showed no significant growth in fidelity score following the initial 12-month measurement. Researchers suggested specific actions by leaders to support the increase in implementation. Those resources included follow-up training, increased amounts of feedback, and changes in personnel.

Further research is needed to determine whether the middle school will continue to see implementation gains. However, the data collected in this case study clearly indicates that while implementation has occurred to some degree across the site, it has not approached an ideal of 100% or the rubric-based threshold of 75%.

Success of Implementation by Predictor Variable

As described in the previous chapters, the fidelity indices were analyzed as a function of years of teaching experience, teachers' primary subject areas, and latency since training.

Relationship between fidelity and years of teaching experience. As can be seen in table 4, as years of experience increased, the overall level of fidelity increased, as well as the fidelity of four of the five critical components (the exception being OLP). The only statistically significant relationship was in the implementation of IRNT. The statistical significance of the relationship between IRNT and years of experience is likely due to chance rather than any noteworthy difference in the component. However, the overall trends warrant further investigation. The trend, though not statistically significant, is supported by data from the teachers' surveys, where all respondents had more than five years of teaching experience. In 100% of the surveys, the respondents indicated that they felt they would be able to fully implement each component within two years. If the respondents are correct, they would have a minimum of seven years of experience when critical components are fully implemented. Unfortunately, due to the low survey response rate, survey data regarding perceptions of teachers with less experience are not available. However, one teacher was interviewed who had less than three years of teaching experience. She explained her level of implementation fidelity: "It's a better way of teaching, but that doesn't mean it's easy. I just don't have the experience to see how it all fits together."

The literature suggests that teachers with limited experience report that emotional exhaustion and pressures from work-related tasks limit their ability to implement changes in practice (Kwakman, 2003; Skaalvik & Skaalvik, 2010). Kwakman (2003) described the inverse relationship between teacher stress and participation in professional development and associated feedback. She noted that perceived stress was a predictor of participation more often than other factors, such as relevance and quality of professional development. CM survey data is not available from teachers with low levels of experience, due to the overall low response rate. However, interview data did provide some insight as to why teachers with limited experience seemed to implement at a lower level. Teachers frequently described a theme that teachers with less experience are overwhelmed by the demands of the profession. Among other challenges, newer teachers feel pressured to use a number of different strategies. This seems to result in less experienced teachers dedicating finite blocks of time to specific tasks or initiatives separately. In contrast, as described by one of the teachers quoted above, more experienced teachers have the ability to see how pieces of different strategies fit together and can therefore “work on them” simultaneously. Thus, the addition of CM strategies did not appear to add to the overall stress of the experienced teachers interviewed.

There was not a universal agreement that years of teaching experience result in increased implementation fidelity. In particular, during interview segments that focused on negative perceptions, teachers were able to describe examples where more teaching experience may hinder CM implementation. For example, when discussing the overall implementation of writing in math classes, a teacher commented:

> Someone will say something like "Well, you're a math teacher. You don't need to be teaching language." And I think, because mostly they're older than me . . . I'm a pretty young teacher, so I think that I don't really ever state my opinion, and I'll listen to what they say. . . . They've had way more years of experience. I definitely respect their experiences, but I would say now in year two, having things calm down a little bit more, math is a language in and of itself. By teaching Constructing Meaning or teaching these sentence frames, you're still, in some way, teaching logical reasoning, which is what you're supposed to be doing in math. For me, in every content area, we need to be teaching language, but that's just me.

This teacher described alignment to the overall principles of CM, but was not engaged with her team to collaborate on implementation. Communication with peers was shown to be a statistically significant challenge for novice teachers in a mixed-methods study of 86 novice (less than two years of experience) teachers (Fantilli & McDougall, 2009). Respondents described the challenge as isolation from peers and a perceived lack of respect from more experienced teachers. Data from this case study indicate similar perceptions by less experienced teachers and may have contributed to the variation in implementation of the components.

Relationship between fidelity and latency since training. As can be seen in table 6, the analysis of data by training cohort did not reveal any noticeable trends or statistically significant differences. There was little variation within components or in the overall fidelity index.

The lack of a difference in cohort implementation does seem to contradict the qualitative findings. For example, 100% of the teachers responding to the survey felt that they should be able to implement each component within two years. If one were to assume based on that sentiment that implementation follows some sort of progression, then it would follow that the earliest-trained teachers would show the greatest level of implementation. That assumption was not borne out in observation.

Relationship between fidelity and teachers' primary subject area. In Chapter II, CM training was described as an intervention to increase the academic language development instruction across the school, in every classroom. The explicit teaching of language in all classes is generally thought to be a shift in practice, particularly in non-writing courses. Based on that shift, there was an expectation that courses not traditionally considered to include writing would show the lowest level of implementation fidelity.

The quantitative results are somewhat in line with the expectation described above. Humanities teachers, who are described as teaching English language arts and/or social studies, generally displayed the highest rates of implementation. One unexpected example, however, was the math department, whose teachers showed implementation near the median level. Math may be considered to traditionally include the least amount of language instruction, above only physical education. However, the sample size of this case study prohibits strong conclusions regarding the relationship between subject area and level of implementation.

RQ2: Conditions that Influence Implementation

As described in chapter I, the main goal of research question 2was to understand the conditions that support or hinder the implementation of CM at a local middle school. The conditions described in this section were derived from the analysis of the qualitative data collected in this case study. Teachers were asked to provide their perceptions regarding what influenced the implementation of CM practices. The section below describes the teachers' perceptions of the training itself, the conditions either in place or desired that can support implementation, and lastly, the conditions that were in place that were perceived to hinder implementation.

Teachers' perceptions of CM training. Teachers' "buy-in" of educational reform initiatives has been considered critical to the implementation of new programs (Datnow & Castellano, 2000; Fullan, 1991; Gulumhussein, 2013; Opfer, 2011). Fullan (1991) suggests teacher buy-in is influenced by whether their beliefs align with the priorities of the initiative. The teachers' overall perception of CM training is a condition that influenced implementation. Teachers' comments regarding the training itself were all generally positive. Although the principal investigator had developed an initial code to organize any negative comments, the analysis of the interviews did not uncover a single quotation referring to anything negative about the training itself. Teachers used these terms and phrases to describe their perceptions of the training: "practical," "best practices," "what we know we should be doing," and "it helps all students, not just ELLs." All teachers interviewed felt that, when fully implemented, the training would lead to gains in student achievement. Survey data yielded similar results, with 100% of respondents feeling that CM practices will have a positive effect on student achievement. Likewise, only 15% of respondents felt that the school should have different priorities for their professional development. It is logical to conclude that the content of the trainings was perceived as positive by the participants in the study.

The materials themselves began to emerge as a theme regarding the perception of the training. During the course of teacher interviews, the teachers spoke of more tangible resources to understand whether there were resources needed that could simply be purchased. Eight of the nine teachers referred to some sort of preprinted resource available directly from E. L. Achieve. One example is the “CM flipbook” and desktop guide that offers teachers examples of immediate use of strategies aligned to each of the critical components. As one teacher mentioned, “Many of the materials provided by CM are easily modified to use almost immediately in class.”

The presenters themselves also contributed to the positive perception of the CM training. One teacher described the trainers as energetic, reporting that the trainers “immediately earned [the teachers’] respect through their knowledge and expertise.” Furthermore, the trainers seemed to stay connected to their trainees beyond the three official days of training. One teacher in an interview named a trainer who “made herself very available for support and help” above and beyond the approachability of the other instructors. None of the teachers I interviewed or surveyed indicated any reluctance or hesitation to contact the trainers for guidance.

Collaboration. Teacher collaboration has been described as a systematic process to analyze and improve instructional practices (Dufour, 2004). As can be seen in table 12, every teacher interviewed spoke of the need to work with others. They used words and phrases such as “collaboration,” “planning,” “working together,” and “share the work” when talking about completing their tasks. The tasks they referred to were all related in some way to the development of CM practices. However, they expressed a need to collaborate with distinctly different groups of people. The groups were peers with similar jobs, or other teachers that taught the same course or subject area; interdisciplinary groups or teachers that taught different subjects but to the same group of students, such as grade-level teams; and collaboration with CM trainers. Based on the structure of professional collaboration in place at the middle school, collaboration within subject groups focused on the development or modification of curriculum and materials to align with CM practices. Likewise, teachers often spoke of dividing tasks and sharing resources. For example, one teacher may be writing the sentence frames for the lesson while another teacher is developing discussion prompts. Similarly, another teacher remarked,

> We work really well together in planning activities that are going to use the vocabulary and get kids talking about what we're doing. That's a huge part of what makes it successful is that we have the opportunity. I'm not trying to do it all by myself.

One teacher described this process during teacher interviews as one that would be beneficial but not currently in practice: “I'd like some time to collaborate with my science partners and work on lesson plans that have CM implemented in them.”

In contrast, collaboration in interdisciplinary teams centered around shared groups of students. Teachers spoke of discussing the rate of gradual release for specific groups of students or strategies that may have worked particularly well for an individual teacher. One teacher commented on the desire to “collaborate even with my other teammates, so my humanities teacher, my science teacher. It would be nice to know what they're seeing, if they're seeing the same patterns with language with our kids."

Collaboration with peers was found to be one of the key components described by Joyce & Showers (2002) of effective professional development. Peer collaboration, as described by their study, included both the development of curricular materials as well as the logistical planning for implementation. Gulumhussein (2013) describes peer collaboration to be the condition in which teachers can apply educational research and develop innovative changes to their practice. In the case study at hand, school leaders should further develop structures to support the peer collaboration between teachers in the middle school.

In addition to peer collaboration, teachers also spoke of the need to collaborate with CM trainers. Specifically, teachers were looking for specific feedback and coaching on their curricula and instruction. Teachers described the need for collaboration with CM trainers as a way to validate their own perceptions of their implementation.

As described earlier in this chapter, teachers consistently over-reported compared to the observations. The discrepancy is meaningful, particularly when considering the implementation by individual teachers. If an individual teacher were to rely only on his or her self-evaluation, the perceived need to adapt and change practice may be less than if he or she received direct feedback from an external observer. When asked about the discrepancy between the reflections and observations, responses included the need for an impartial observer to give direct feedback. For example, one teacher described the difference between being observed by a colleague and a trainer in this way:

> There's the past, the history, the knowing of each other, and that might not be as fruitful as it could be if it was just someone from the outside who's not in charge

> of evaluating me, they're not associated with, they're not a friend of mine, or someone that I used to teach with, or a colleague. It's just someone from CM, whom I don't know, coming in, and watching me use the strategies and giving me some really effective feedback and coaching.

Feedback by trained coaches, as opposed to peers, may be advantageous (Scheeler, Ruhl, & McAfee, 2004). Trained coaches simply have more experience with specific interventions and are able to provide more specific feedback to the associated practices than peer coaches who may still be learning the programs themselves (Mallette, Maheady, & Harper, 1999). In contrast, however, trained coaches may not have informal access to teachers, and their observations are more likely to be scheduled and less frequent than visits from internal coaches (Nishimura, 2014). Implementation plans should include opportunities for regular feedback for the teachers. Practically, the feedback plan will need to include peers, due to resource limitations. However, feedback from experts should not omitted.

Time as a resource. The concept of time as a resource was a complex theme during this case study. In 100% of interviews, the need for additional time to implement the CM practices was mentioned. However, in some cases, time simply predicated a need for a different resource, indicating that it was the latter that was actually in need. For example, one teacher stated: “I am a science teacher, not a language arts teacher. I need time to learn how teach this way.” Although the term *time* is used, the comment actually reveals a need for additional training and support, as described in the preceding section, rather than simply more time.

In contrast, other respondents were able to make statements indicating that time itself was the resource needed. Seven of the nine interviews mentioned the need for more time with the students. A characteristic response was, “These strategies are better, but they certainly take more time than just lecturing and moving on." Interestingly, during the case study, teachers were transitioning from 90-minute instruction period blocks to 75-minute blocks. Therefore, it is possible that the perception of a need for more time with students was due to the change in the schedule and not the addition of CM practices. In addition, five of the nine teachers interviewed discussed simply needing time to modify their existing curricula to meet the requirements of CM practices. One teacher commented that when time was provided to plan with the team, implementation "was done with ease."

Although a specific guideline for the amount of time needed for workshops, coaching, and follow-up is not agreed upon in the literature, the concept of implementation timelines has been addressed in various studies. In her report to the National Staff Development Council, Linda Darling-Hammond and colleagues (2009) analyzed nine experimental design studies measuring the relationship between time and implementation of PD. The researchers found that in every study, the duration of the training (including coaching and follow-up) was positively associated with implementation fidelity. The training time needed may be much higher than what is provided in traditional workshops that last one to five days. In various reports on professional development, between 50 and 80 hours of direct engagement is needed to significantly influence teacher practices (Corcoran, McVay, & Riordan, 2003; Supovitz & Turner, 2000; Wagner & French, 2010).

The teachers in this study each completed two full days of direct training (approximately 16 hours). Additional time (up to eight hours per teacher) was provided for teachers to opt to spend time with trainers adapting their curricular materials. Beyond the formal time provided, teachers have described using planning and personal time to implement the strategies. Although teachers were not asked to quantify the total number of hours they have spent working on implementation, it does not seem likely that many, if any, have approached the minimum 50 hours suggested above.

An interesting discrepancy was presented from the survey data in regards to time as condition of implementation. Only 8% of respondents stated that CM practices are too time-consuming to implement. However, it is less clear whether the school providing more time would aid implementation. During the surveys, 76% of teachers reported that they were able to collaborate with their "learning team" regarding CM practices. Learning teams are district mandates, similar to the structure of professional learning communities described by Dufour and Dufour (2002). According to the collective bargaining agreement, learning teams are to be allocated a minimum of 90 minutes a month, during the contract hours, to meet and collaborate. However, on a subsequent question, only 39% of respondents agreed that there is time available in the school day to plan for CM implementation. The survey data suggests that there may be conflicting understandings as to how the time that is available to teachers is to be used, and whether implementation of CM practices should take priority during that time.

It is clear that teachers will need to invest substantial amounts of time in order to approach the necessary investments required for full implementation. Under the current schedule utilized by the school, the collaboration will need to occur during preparatory periods, after school contracted time, or during one of the three staff development days that occur throughout the year.

Alignment with Other Priorities

As described earlier in this chapter, the evidence collected in this case study suggests that there are areas of overlap between CM practices and other district priorities. CM training was selected explicitly after an emphasis on SIOP training. Teachers and administrators felt that the ideals and philosophies described in SIOP training were operationalized by CM practices. Additionally, the school, and its associated district, are heavily invested in the implementation of AVID, a structure and curriculum aimed at increasing college success in traditionally underrepresented groups. One of the key components of AVID is the teachers' use of Writing, Inquiry, Collaboration, Organization, and Reading (WICOR) strategies. The teachers that were interviewed were able to see the similarity between CM practices and WICOR strategies. These connections were described by teachers as positive and contributed to the feeling that CM is not "just another thing." For example, one veteran teacher commented, "what really connected was the strategies with AVID, and then partnering that with Constructing Meaning. It was like a really good additional underpinning for the strategies in AVID." Another teacher said that CM "provided more structure and specifics to techniques, handouts, etc., that I had been working to implement as a result of my work with SIOP and more recently AVID trainings."

Alignment with other priorities as a hindrance to implementation. The connections to other initiatives did include some negative remarks. The remarks did not suggest that any one initiative is negative, but that they each place a drain on resources, and teachers said they felt like there were some competing priorities. As one respondent stated, "AVID is great, but it is taking away from the emphasis of CM, like we aren't really doing CM anymore.” Another teacher simply said, “My focus has also been on AVID much more than CM.”

In a similar explanatory study that investigated the implementation of Positive Behavior Intervention and Supports (PBIS), 53% percent of teachers surveyed indicated that explicit prioritization of the program aided in implementation (Andreou, McIntosh, Ross, & Kahn, 2015). Teachers stated that the continued prioritization, over multiple years, helped to “validate the program” and increased the likelihood that they would “alter their practice” (p. 164).

Limitations of the Study

The discussion of the analysis of research questions presented is intended to be helpful to school stakeholders as they evaluate and plan for their ongoing implementation of CM practices. Guidance has been provided to school district leaders as they adopt and plan professional development activities for their staff. However, the methods described in this case study were not without limitations. The limitations and their potential impacts on the findings of this study are discussed below.

Psychometric properties of the rubric. The *Refining Our Practices Rubric* was at the center of all quantitative analysis in this case study. The lack of psychometric data for the rubric calls into question the accuracy of the findings. Although this study did include a reliability analysis of both the raters and the items within each component, the data used came from a relatively small number of observations, n = 5 and n = 30, respectively. The reliability analysis did provide useful information, as described in Chapter IV. For example, in the case of the item analysis, item AWS(3) showed limited agreement with other AWS items and was removed from analysis. However, a larger scale reliability analysis is needed prior to removing the item from future uses of the rubric. Additionally, the agreement between the two observers only met minimally acceptable levels of reliability (kappa = .607). Therefore, the variance in the data presented may be more a function of rater disagreement than of differences in actual implementation fidelity. However, it should be noted that the rubric allowed for a range of four possible scores. The application of the kappa statisitic did not likewise account for a range of agreement. Rather, the calculation treated the disgareement between scores of one and four on the same item as equivalent to the difference between a three and a four. A simple difference analysis indicated that approximately 70% of the measurements that were in disagreement between the two raters were within one level on the rubric. The small range in disagreement provides confidence about the reliability of the observational data not completely reflected in the calculation of the kappa statistic.

The evaluation was formative in nature, meaning it was intended to provide insight to the school as to how to improve the implementaiton process of CM training. There was no intention to use any of the data to make summative judgements on the continuation of the program, and certainly not to make any job performance claims about any individual or the school in general. However, the lack of substantial agreement was a point of concern and is further addressed below.

The validity of the rubric was assumed, which presented a significant limitation of the study. The items in the rubric called for a very narrow range of observations for each of the components. However, the observations required for each item may not detail a complete representation of the construct presented by the component. Therefore, a teacher may be implementing practices that are in line with a particular component that are not specifically included in the rubric. If that were the case, the teacher would receive an innappropriatly lower score. Further analysis on the rubric is needed to determine its appropriateness and robustness as an evaluation tool, discussed below.

Sample size of the study. All statistical analyses were completed without sufficient statistical power, due to the small sample size. As a result, the findings presented here should be further investigated through studies with greater participation prior to making summative claims. The findings in this study presented school leaders with some potentially useful information for planning, but more investigation is needed before making high-stakes changes to the program. Furthermore, the study was designed and conducted completely within the context of the specific school, and findings are not generalizable to other CM implementation contexts or other teacher PD programs.

Participant bias. Participation in this study was voluntary. During the observation phase, only four teachers "opted out." However, during the survey and reflection phase, only 13 of the 30 teachers participated. An attempt to uncover the reasons behind the lack of participation yielded little conclusive information. The only feedback provided was that time was too valuable to spend on nonessential tasks.

While nonresponsive teachers cited only time constraints, there may be specific groups of teachers whose perceptions were not represented in surveys and reflections. The limited range of teacher perceptions should be noted when considering possible bias in this case study. For example, of the teachers who responded to reflections and surveys, none of them were within their first five years of teaching. As described above, teachers in the first few years of teaching describe feeling exhaustion and burnout from pressures at work (Fives, Hamman, & Olivarez, 2007; Kwakman, 2003). If the teachers in this study also feel exhausted during their first years of teaching, they may have had more negative perceptions of CM practices than the more experienced teachers who responded (generally positively) to the survey. There is no way to know the motivations of the teachers who did or did not respond to surveys and reflections. Those with a positive impression of the program might be more likely to respond, and less likely to voice challenges or hindrances to implementation.

Over-reporting by teachers. Within the participant bias of the study was the over-reporting of implementation by teachers. Over-reporting emerged as a theme throughout the quantitative analysis of this study. In the case of every component, except UBD and the overall index, teachers' self-reported scores were higher than those reported from direct observation by the coaches. The differences were found to be statistically significant in AWS, IRNT, and OLP. The higher reporting was a topic discussed in every interview. No teacher made any statements of surprise upon learning that teachers had reported implementation higher than the observers, and many were able to offer opinions as to the reasons.

Two common opinions regarding the over-reporting emerged. The first opinion was that teachers may be reporting on what they *planned* to do or could *describe* doing, without actually taking into account whether the action had actually taken place. The second opinion is that the over-reporting indicates a need for direct observation by and feedback from trained CM coaches. Several teachers addressed these ideas during the interviews. One respondent said, “I found myself being like, ‘Yeah, I do that,’ and then I overestimate how much I do that. In reality, [the students are] not having as many conversations as I think they're having.” Another teacher commented, “I think that we want to look like we know what we're doing, for one thing. Put a rosy spin on it. Also, we might have an idea that we're thinking of . . . in our head." Further remarks on this issue include the following quotes from two teachers:

> We have it in our head. . . . One of my favorite lines is, "I have the best lesson plans in the world. It's going to be awesome." Then the kids show up. They don't do what I'm thinking. "Oh, they'll do this, they'll do that." That's some of it.
>
> In my head, as I develop those lessons or want to use these [strategies], I'm thinking about these. I'm not really making it happen in the classroom as effectively as I am thinking about it. I think I know how to use this. I think that I'm using it, but it's not really coming across.

The concept that teachers, or any practitioner, would over-report is not surprising and has been well documented in the literature (Eva & Regehr, 2008; Kruger & Dunning, 1999). Described by Kruger and Dunning as "unskilled and unaware" (1999), there is general consensus that individuals have a poor ability to self-assess accurately. However, self-assessments have been shown to have positive contributions to the implementation and evaluation of professional development (Eva & Regehr, 2005; Langendyk, 2006). Self-assessments provide opportunities for individuals to describe the contextual conditions affecting PD and to reflect on their own progress toward implementation. Although the progress described is likely inflated, the opportunity to reflect provides teachers with the opportunity to better understand their practices. In the case study presented here, the over-reporting by teachers suggests a need for more observation by trained observers in order to judge the actual level of implementation and give support where necessary. This quote summarizes the need:

> I imagine that when you feel like you're working really hard to do something new—doing new things and incorporating new things in your practice is difficult, and if you feel like you're working really hard at it, even if you're not doing a good job at it, you are maybe rating yourself in terms of how hard you *feel* like you're working at it versus how well you're doing, which is why I think having

> someone observe and give feedback, and do coaching is a really vital part of any professional development. And it's the most absent part of professional development.

Target of implementation. RQ1 investigated the success of implementation. To compound the measurement challenges presented by the rubric and small sample sizes, there was also a lack of a predetermined standard. There were no specific success criteria described during the training or set by the administrators of the school. The targets used as a standard in this study were developed in consultation with district leaders and CM trainers, and included a review of the relevant literature. However, it must be noted that the standard was set for the purposes of this study and evaluation only. The standard of addressing success by investigating the implementation of the components, the variance in implementation, and a threshold based on the rubric were useful in making recommendations to the school. However, the lack of a predetermined standard during the initial design and training is a topic for consideration presented as an application for future research.

Implications for Practice

In the following sections, the interpretation of the results of this case study and the associated implications for the middle school and its parent district are presented. The recommendations presented are related to determining a target for successful implementation, developing reliable systems of observation, factoring in time as a resource, integrating with other priorities, and distinguishing between training and intervention. Recommendations for the general research community are also presented.

Target for Successful Implementation

As discussed above, success criteria needed to be established for this study. The lack of a clearly defined implementation plan has been cited as a hindrance in programs that have failed to be adequately implemented (McGrew et al., 1994). Schools lacking a clear plan have noted that interventions seem to simply fade away from their practice over time (Andreou et al., 2015). The school in this study had not determined expected levels of implementation over time. In fact, there was little to no specificity provided to the expectations from management in regards to which instructional practices should be utilized. Rather, the general statement of "we should see these practices in every room" was the only basis for successful implementation, prior to this study. Likewise, E. L. Achieve does not provide an implementation timeline, and neither the district nor the school chose to develop one prior to the training. During the case study, it became apparent that the approach of school and district leadership was to celebrate the areas that were implemented rather than focusing on the areas that were not.

It is not too late for the school to develop such a plan. One application of this case study could be to use the indices of the components as a baseline and to set goals for expected future gains. School leaders could develop a systematic plan that addresses the conditions needed for implementation as well as the challenges of evaluation that have been described in this study. Teachers would be able to receive the specific feedback they need, based on the goals of the implementation plan. As discussed above, teachers felt that observations and feedback on specific CM components, made by CM experts, would better support successful implementation. With clear, component-specific plans in place, feedback could be highly targeted.

Developing Reliable Systems of Observation

The continued implementation of CM practices by the school should include the continuous evaluation of classroom practices. As described above, the teacher comments collected in this case study indicated that observations and feedback supported implementation. Both the data collected and the relevant literature indicate benefits of both peer and outside (CM trainer) observation and feedback cycles. The school should design a program where teachers observe and provide regular feedback to each other as well as periodically bringing in trained CM coaches. The practice of observations and feedback will support teachers in their implementation and provide leaders with evaluative data that can be used to make program adjustments. The methods for the evaluation of CM practices by utilizing both quantitative and qualitative methods, as utilized in this case study, could be repeated, if the institutional limitations discussed above are addressed.

Increased reliability analysis of the *Refining Our Practices Rubric*. The reliability analysis that was presented in this study was derived from data collected from a very small sample. The lack of confirmed psychometric data on the tool needs to be addressed to support large-scale evaluations within the school and the greater district. Ideally, the data would come from E. L. Achieve, which has supported CM schools across the country. However, the district has trained enough teachers to produce a sample size adequate for such analysis. As explained in Chapter III, the goal of the district was to have every middle school teacher trained in CM practices. If the district were to achieve this goal, a comprehensive study would contain a sample size of approximately 400 teachers. A systematic analysis of CM implementation could be used to better understand the psychometric properties of the rubric. If the rubric is shown to be reliable and valid, it could be used immediately in studies similar to the one presented here.

Considerations of Time as a Resource

In addition to continued evaluation, school leaders should support the conditions that have been described to support implementation. Teachers' consistency in describing time as a necessary condition to support successful implementation did not come as a surprise. However, the discrepancy between some teachers feeling that adequate time is provided while other teachers felt it was not was interesting. Particularly, the majority of the teachers in this case study were on the same bell schedule with the same number of preparatory minutes. Therefore the discrepancy could not have resulted from simple differences in schedules. Rather, the discrepancy may have resulted from the teachers' prioritization of CM practices and their use of the time that they did have. Once a clear expectation of an implementation timeline is put in place, teachers could be given clear objectives to accomplish during available preparatory time.

Connections to Other Priorities

During this case study, the connection to other school and district initiatives was perceived by some teachers as a support to implementation. In particular, teachers spoke of CM practices as a continuation of Sheltered Instruction Observation Protocol (SIOP) trainings. Generally, they described that connection as positive. Teachers also spoke of CM training as being connected to the training received through Advancement Via Individual Determination (AVID) trainings. The shared strategies, such as interactive note-taking, were seen as complimentary and positive. However, further probing into the connection between AVID and CM indicated that some teachers perceived the connection as a hindrance to CM implementation.

The district has been financially supported by a Fortune 500 corporation to provide training and implementation of the AVID program at all of its schools. While the foci and practices of AVID and CM align, they are two distinct programs with different critical components and measures of evaluation. Multiple teachers felt that they did not have the capacity to implement both programs simultaneously. What has resulted was the feeling that CM was more of a one-time event, where teachers would use what they felt was beneficial and adapt it to their practices. CM practices were not considered as a systematic intervention but rather a workshop that would supplement ongoing instructional techniques. As described in Chapter II, professional development that is delivered in a workshop format is less likely to be fully implemented.

Distinguishing Between Training and Intervention

Interview comments indicated uncertainty as to whether CM practices will serve as a standard for teaching practices across subject areas. In one regard, CM training and its associated practices can be thought of as an intervention, which is the approach used by the school and district at the time of this study. Considering the practices as an intervention implies that the practices will be implemented as intended and with an adequate level of fidelity. However, in light of the implementation of AVID and other school priorities, it is also possible to simply consider CM training as a workshop. Workshops tend to be one-time events that may help to improve the practices of certain teachers but do not result in widespread change or improvements to overall outcomes. School leaders should clearly communicate that full implementation of CM practices is consistently expected of all teachers. School leaders should also allocate the appropriate resources, including time, that are necessary to expect full implementation.

Chapter II outlined the need for fidelity measurement to be included in the implementation plans for interventions. The current study was an applied research project with a primary goal of presenting school-specific findings and recommendations. This study also identifies several areas in need of further study by the broader research community.

This study was able to design and utilize a model based on prior studies and recommendations. In addition to describing a practical application, this study also uncovered challenges that should be addressed by the professional development community. Specifically, challenges to schools completing research-based internal evaluations were uncovered. The substantial challenges uncovered were in the area of valid and reliable measures of implementation, and an established standard for implementation.

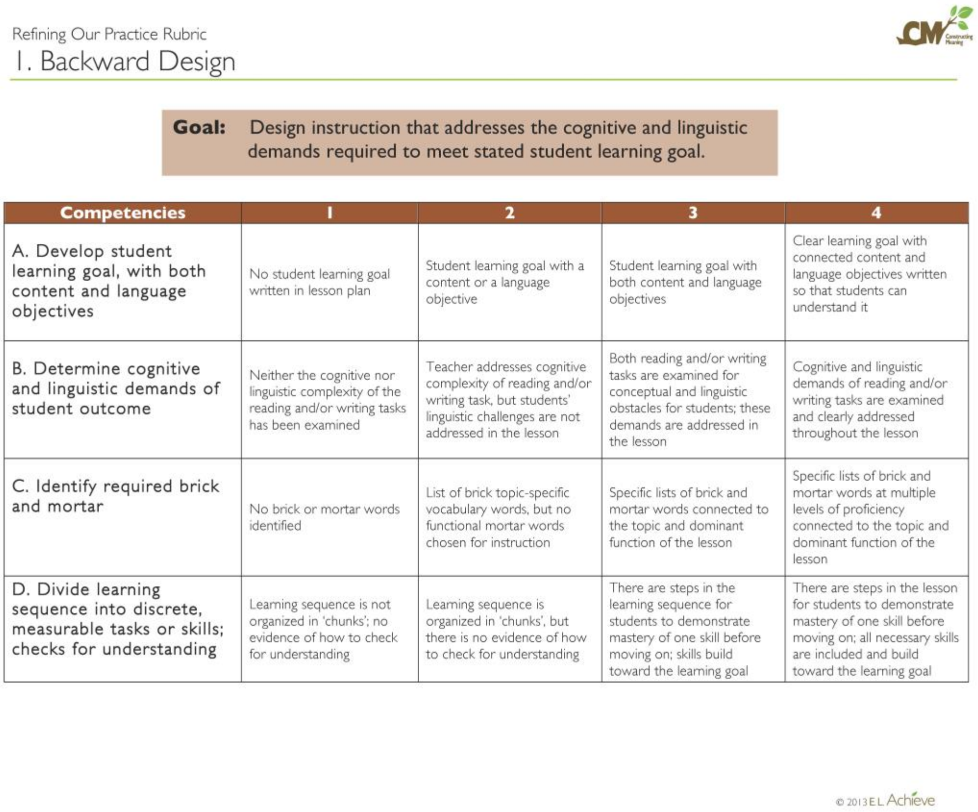
The critical components of CM were readily identified by the developers of the program. However, as described above, the rubric used in the evaluation lacked accompanying psychometric data. School personnel should not be charged with assessing the evaluation tools of interventions they choose to implement. Developers should provide valid instruments as well as clear guidance for measuring and achieving reliability when offering professional development packages to schools or districts. Additionally, PD providers should define standards of implementation, using accepted methods.

Implications for Future Research

This study was conducted to not only measure implementation, but to also *understand* it. Researchers utilizing fidelity studies, either as formative evaluations or within broad studies of program effect, should consider the inclusion of qualitative methods in their analyses, in order to better understand the contextual conditions that influence the implementation of the intervention they are studying. In the current case, the qualitative methods allowed the providers of the intervention to describe perceptions and nuances for the intervention that would be difficult to ascertain through solely quantitative methods. The results of studies that combine both quantitative and qualitative results are thus more likely to be recognizably useful and more likely to be applied by local school leaders. The use of well-designed evaluations of implementation is crucial for school leaders who are attempting to raise student achievement outcomes through quality professional development. The use of implementation data throughout the PD process is crucial. Implementation data will provide school leaders with the information that is necessary for timeline adjustments and resource allocations. Implementation data is also a vital component to showing the efficacy of the programs when included as part of a well-designed experimental study.

APPENDIX

THE REFINING OUR PRACTICES RUBRIC

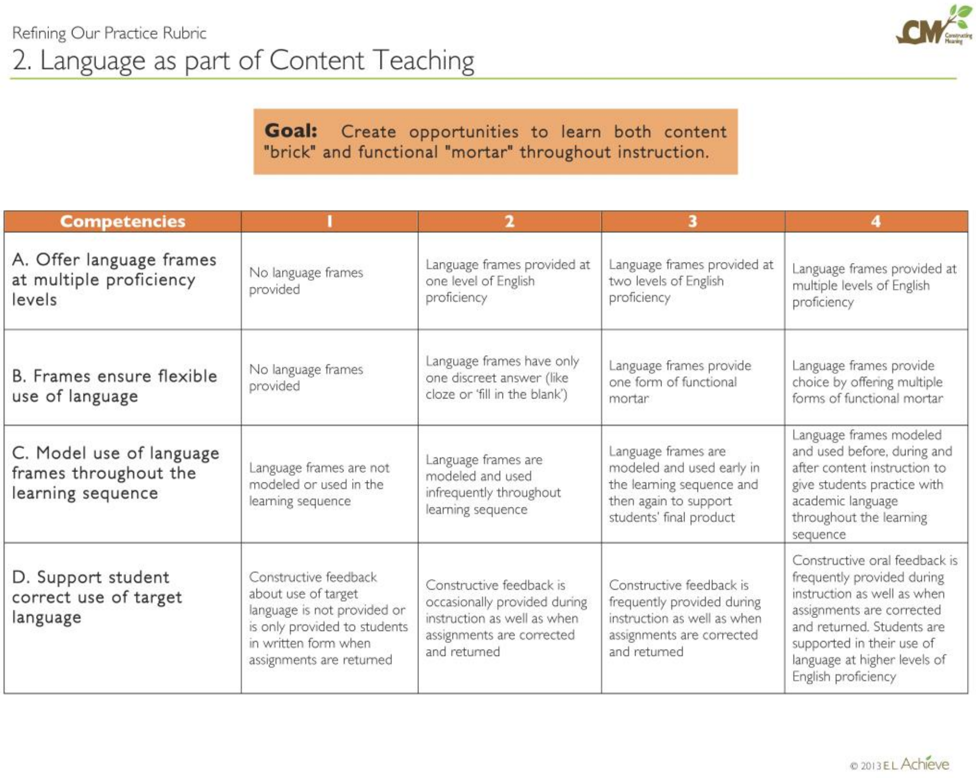


Refining Our Practice Rubric

I. Backward Design

Goal: Design instruction that addresses the cognitive and linguistic demands required to meet stated student learning goal.

Competencies	1	2	3	4
A. Develop student learning goal, with both content and language objectives	No student learning goal written in lesson plan	Student learning goal with a content or a language objective	Student learning goal with both content and language objectives	Clear learning goal with connected content and language objectives written so that students can understand it
B. Determine cognitive and linguistic demands of student outcome	Neither the cognitive nor linguistic complexity of the reading and/or writing tasks has been examined	Teacher addresses cognitive complexity of reading and/or writing task, but students' linguistic challenges are not addressed in the lesson	Both reading and/or writing tasks are examined for conceptual and linguistic obstacles for students; these demands are addressed in the lesson	Cognitive and linguistic demands of reading and/or writing tasks are examined and clearly addressed throughout the lesson
C. Identify required brick and mortar	No brick or mortar words identified	List of brick topic-specific vocabulary words, but no functional mortar words chosen for instruction	Specific lists of brick and mortar words connected to the topic and dominant function of the lesson	Specific lists of brick and mortar words at multiple levels of proficiency connected to the topic and dominant function of the lesson
D. Divide learning sequence into discrete, measurable tasks or skills; checks for understanding	Learning sequence is not organized in 'chunks'; no evidence of how to check for understanding	Learning sequence is organized in 'chunks', but there is no evidence of how to check for understanding	There are steps in the learning sequence for students to demonstrate mastery of one skill before moving on; skills build toward the learning goal	There are steps in the lesson for students to demonstrate mastery of one skill before moving on; all necessary skills are included and build toward the learning goal

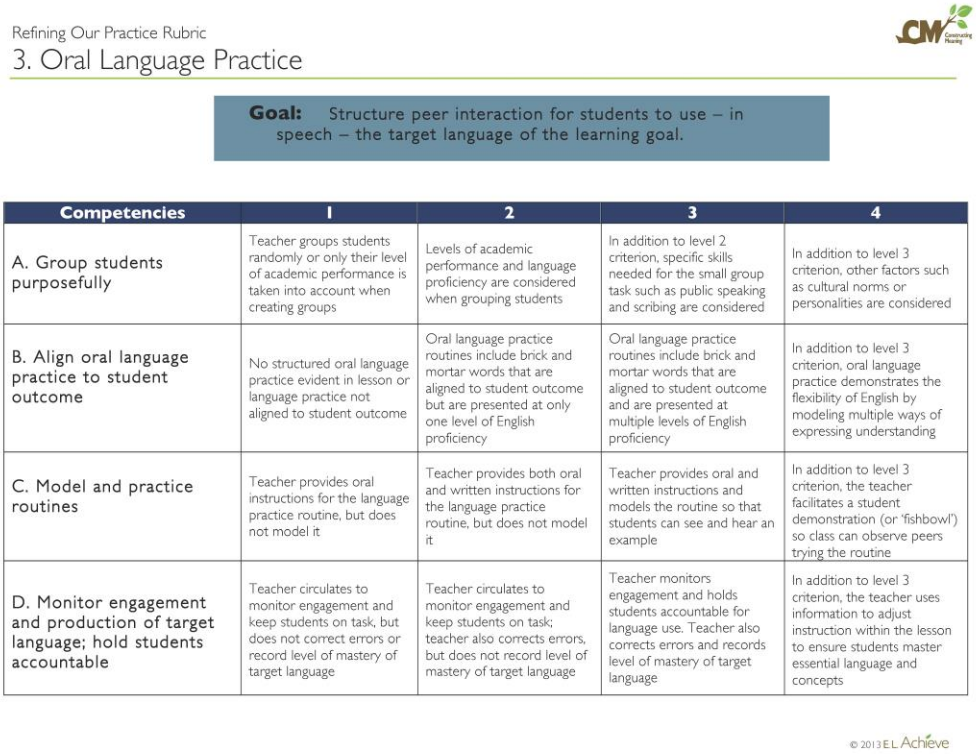


Refining Our Practice Rubric

2. Language as part of Content Teaching

Goal: Create opportunities to learn both content "brick" and functional "mortar" throughout instruction.

Competencies	1	2	3	4
A. Offer language frames at multiple proficiency levels	No language frames provided	Language frames provided at one level of English proficiency	Language frames provided at two levels of English proficiency	Language frames provided at multiple levels of English proficiency
B. Frames ensure flexible use of language	No language frames provided	Language frames have only one discreet answer (like cloze or 'fill in the blank')	Language frames provide one form of functional mortar	Language frames provide choice by offering multiple forms of functional mortar
C. Model use of language frames throughout the learning sequence	Language frames are not modeled or used in the learning sequence	Language frames are modeled and used infrequently throughout learning sequence	Language frames are modeled and used early in the learning sequence and then again to support students' final product	Language frames modeled and used before, during and after content instruction to give students practice with academic language throughout the learning sequence
D. Support student correct use of target language	Constructive feedback about use of target language is not provided or is only provided to students in written form when assignments are returned	Constructive feedback is occasionally provided during instruction as well as when assignments are corrected and returned	Constructive feedback is frequently provided during instruction as well as when assignments are corrected and returned	Constructive oral feedback is frequently provided during instruction as well as when assignments are corrected and returned. Students are supported in their use of language at higher levels of English proficiency

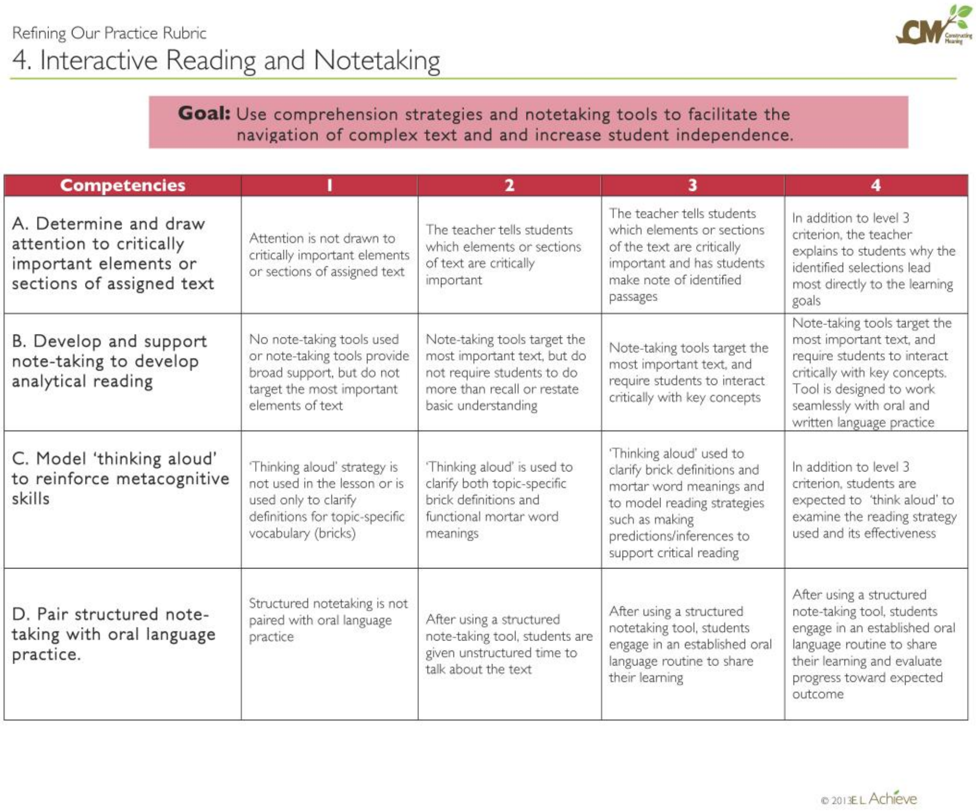


Refining Our Practice Rubric

3. Oral Language Practice

Goal: Structure peer interaction for students to use – in speech – the target language of the learning goal.

Competencies	1	2	3	4
A. Group students purposefully	Teacher groups students randomly or only their level of academic performance is taken into account when creating groups	Levels of academic performance and language proficiency are considered when grouping students	In addition to level 2 criterion, specific skills needed for the small group task such as public speaking and scribing are considered	In addition to level 3 criterion, other factors such as cultural norms or personalities are considered
B. Align oral language practice to student outcome	No structured oral language practice evident in lesson or language practice not aligned to student outcome	Oral language practice routines include brick and mortar words that are aligned to student outcome but are presented at only one level of English proficiency	Oral language practice routines include brick and mortar words that are aligned to student outcome and are presented at multiple levels of English proficiency	In addition to level 3 criterion, oral language practice demonstrates the flexibility of English by modeling multiple ways of expressing understanding
C. Model and practice routines	Teacher provides oral instructions for the language practice routine, but does not model it	Teacher provides both oral and written instructions for the language practice routine, but does not model it	Teacher provides oral and written instructions and models the routine so that students can see and hear an example	In addition to level 3 criterion, the teacher facilitates a student demonstration (or 'fishbowl') so class can observe peers trying the routine
D. Monitor engagement and production of target language; hold students accountable	Teacher circulates to monitor engagement and keep students on task, but does not correct errors or record level of mastery of target language	Teacher circulates to monitor engagement and keep students on task; teacher also corrects errors, but does not record level of mastery of target language	Teacher monitors engagement and holds students accountable for language use. Teacher also corrects errors and records level of mastery of target language	In addition to level 3 criterion, the teacher uses information to adjust instruction within the lesson to ensure students master essential language and concepts

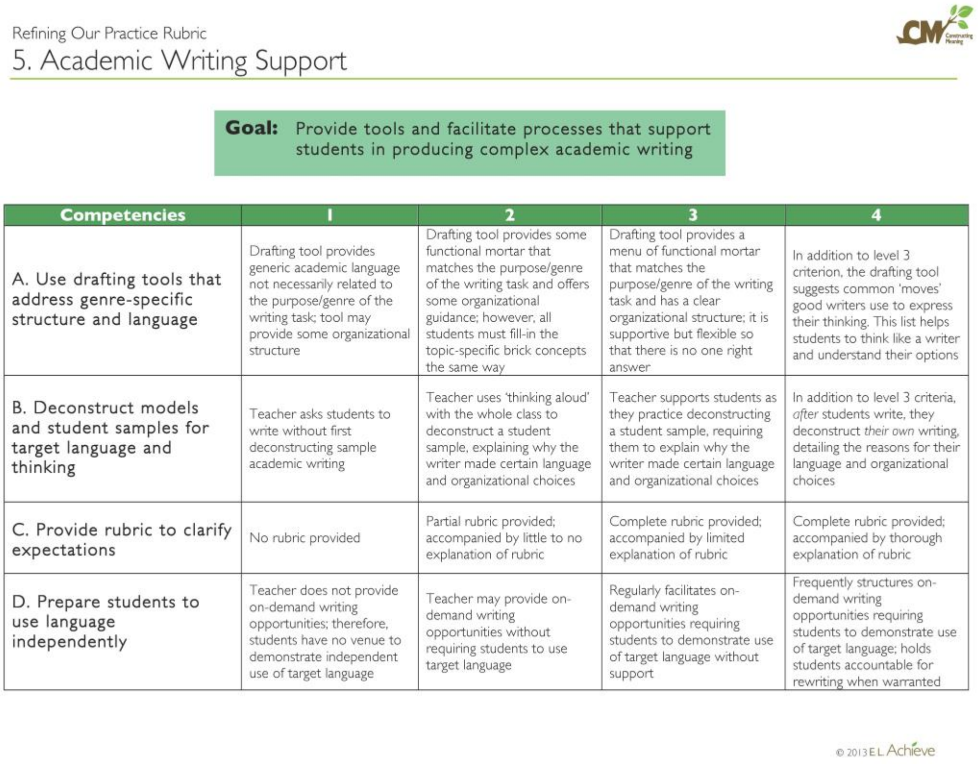


Refining Our Practice Rubric

4. Interactive Reading and Notetaking

Goal: Use comprehension strategies and notetaking tools to facilitate the navigation of complex text and and increase student independence.

Competencies	1	2	3	4
A. Determine and draw attention to critically important elements or sections of assigned text	Attention is not drawn to critically important elements or sections of assigned text	The teacher tells students which elements or sections of text are critically important	The teacher tells students which elements or sections of the text are critically important and has students make note of identified passages	In addition to level 3 criterion, the teacher explains to students why the identified selections lead most directly to the learning goals
B. Develop and support note-taking to develop analytical reading	No note-taking tools used or note-taking tools provide broad support, but do not target the most important elements of text	Note-taking tools target the most important text, but do not require students to do more than recall or restate basic understanding	Note-taking tools target the most important text, and require students to interact critically with key concepts	Note-taking tools target the most important text, and require students to interact critically with key concepts. Tool is designed to work seamlessly with oral and written language practice
C. Model 'thinking aloud' to reinforce metacognitive skills	'Thinking aloud' strategy is not used in the lesson or is used only to clarify definitions for topic-specific vocabulary (bricks)	'Thinking aloud' is used to clarify both topic-specific brick definitions and functional mortar word meanings	'Thinking aloud' used to clarify brick definitions and mortar word meanings and to model reading strategies such as making predictions/inferences to support critical reading	In addition to level 3 criterion, students are expected to 'think aloud' to examine the reading strategy used and its effectiveness
D. Pair structured note-taking with oral language practice.	Structured notetaking is not paired with oral language practice	After using a structured note-taking tool, students are given unstructured time to talk about the text	After using a structured notetaking tool, students engage in an established oral language routine to share their learning	After using a structured note-taking tool, students engage in an established oral language routine to share their learning and evaluate progress toward expected outcome



Refining Our Practice Rubric

5. Academic Writing Support

Goal: Provide tools and facilitate processes that support students in producing complex academic writing

Competencies	1	2	3	4
A. Use drafting tools that address genre-specific structure and language	Drafting tool provides generic academic language not necessarily related to the purpose/genre of the writing task; tool may provide some organizational structure	Drafting tool provides some functional mortar that matches the purpose/genre of the writing task and offers some organizational guidance; however, all students must fill-in the topic-specific brick concepts the same way	Drafting tool provides a menu of functional mortar that matches the purpose/genre of the writing task and has a clear organizational structure; it is supportive but flexible so that there is no one right answer	In addition to level 3 criterion, the drafting tool suggests common 'moves' good writers use to express their thinking. This list helps students to think like a writer and understand their options
B. Deconstruct models and student samples for target language and thinking	Teacher asks students to write without first deconstructing sample academic writing	Teacher uses 'thinking aloud' with the whole class to deconstruct a student sample, explaining why the writer made certain language and organizational choices	Teacher supports students as they practice deconstructing a student sample, requiring them to explain why the writer made certain language and organizational choices	In addition to level 3 criteria, *after* students write, they deconstruct *their own* writing, detailing the reasons for their language and organizational choices
C. Provide rubric to clarify expectations	No rubric provided	Partial rubric provided; accompanied by little to no explanation of rubric	Complete rubric provided; accompanied by limited explanation of rubric	Complete rubric provided; accompanied by thorough explanation of rubric
D. Prepare students to use language independently	Teacher does not provide on-demand writing opportunities; therefore, students have no venue to demonstrate independent use of target language	Teacher may provide on-demand writing opportunities without requiring students to use target language	Regularly facilitates on-demand writing opportunities requiring students to demonstrate use of target language without support	Frequently structures on-demand writing opportunities requiring students to demonstrate use of target language; holds students accountable for rewriting when warranted

REFERENCES CITED

Andreou, T. E., McIntosh, K., Ross, S. W., & Kahn, J. D. (2015). Critical incidents in sustaining School-Wide Positive Behavioral Interventions and Supports. *The Journal of Special Education*, *49*, 157–167. doi:10.1177/0022466914554298

Archer, A., & Hughes, C. (2011). *Explicit instruction: Effective and efficient teaching*. *Explicit instruction: Effective and efficient teaching* (1st ed.). New York, NY: The Guilford Press.

Babbie, E. (2007). *The Practice of Social Research* (11th ed.). Belmont, CA: Thomsen Wadsworth.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*, 29–58. doi:10.3102/00346543074001029

Benner, G. J., Nelson, J. R., Stage, S. A., & Ralston, N. C. (2011). The influence of fidelity of implementation on the reading outcomes of middle school students experiencing reading difficulties. *Remedial and Special Education*, *32*, 79–88. doi:10.1177/0741932510361265

Bennett, S. M., & Hart, S. M. (2015). Reading horizons addressing the “shift”: Preparing preservice secondary teachers for the common core. *Reading Horizons*, *53*(4), 43–65.

Bickmore, K., & Parker, C. (2014). Constructive conflict talk in classrooms: Divergent approaches to addressing divergent perspectives. *Theory & Research in Social Education*, *42*, 291–335. doi:10.1080/00933104.2014.901199

Blank, R. K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement.* Washington D.C. Retrieved from www.ccsso.org

Bond, G., Williams, J., Evans, L., Salyers, M., Kim, H.-W., Sharpe, H., & Leff, S. H. (2000). Psychiatric Rehabilitation Fidelity Toolkit. Cambridge, MA: *Human Services Research Institute*.

Bradshaw, C. P., Barrett, S., & Bloom, J. (2004). *The implementation phases inventory (IPI)*. Baltimore, MD. Retrieved from http://www.pbismaryland.org/forms.htm

Bradshaw, C. P., Debnam, K., Koth, C. W., & Leaf, P. (2008). Preliminary validation of the Implementation Phases Inventory for assessing fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions*, *11*, 145–160. doi:10.1177/1098300708319126

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10*, 100–115. doi:10.1007/s11121-008-0114-9

Brandon, P. R. (1998). Stakeholder participation for the purpose of helping ensure evaluation validity: Bridging the gap between collaborative and non-collaborative evaluations. *American Journal of Evaluation*, *19,* 325–337. doi:10.1177/109821409801900305

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106–148. doi:10.1111/j.1467-6494.1986.tb00391.x

Brunette, M. F., Asher, D., Whitley, R., Lutz, W. J., Wieder, B. L., Jones, A. M., & McHugo, G. J. (2008). Implementation of integrated dual disorders treatment: A qualitative analysis of facilitators and barriers. *Psychiatric Services (Washington, D.C.), 59*, 989–95. doi:10.1176/appi.ps.59.9.989

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 9,* 1–9. doi:10.1186/1748-5908-2-40

Century, J., Cassata, A., Rudnick, M., & Freeman, C. (2012). Measuring enactment of innovations and the factors that affect implementation and sustainability: Moving toward common language and shared conceptual understanding. *The Journal of Behavioral Health Services & Research*, *39*, 343–61. doi:10.1007/s11414-012-9287-x

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*, 199–218. doi:10.1177/1098214010366173

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

Cizek, G. K., & Bunch, M. B. (2007). *Standard setting: A guide establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publicaitons.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319. doi:10.1037/1040-3590.7.3.309

Codding, R. S., Feinberg, A. B., Dunn, E. K., & Pace, G. M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysis*, *38*, 205–219. doi:10.1901/jaba.2005.98-04

Collins, K. M. T., Onwuegbuzie, A. J., & Sutton, I. L. (2006). A model incorporating the rationale and purpose for conducting mixed-methods research in special education and beyond. *Learning Disabilities*, 4(1), 67–100.

Corcoran, T., McVay, S., & Riordan, K. (2003). *Getting it right: The MISE approach to professional development*. Philidelphia, PA. Retrieved from https://cpre.org/images/stories/cpre_pdfs/rr55.pdf

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Los Angeles, CA: Sage Publications.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Dane, A. V, & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9455622

Darling-Hammond, L., & Wei, R. C. (2009). *Professional learning in the learning profession : A status report on teacher development in the united states and abroad*. Washington, DC: National Staff Development Council.

Datnow, A., & Castellano, M. (2000). Teachers' responses to Success for All: How beliefs, experiences, and adaptations shape implementation. *American Educational Research Journal*, *37*, 775–799. doi:10.3102/00028312037003775

DeWitt, J., & Hohenstein, J. (2010). School trips and classroom lessons: An investigation into teacher-student talk in two settings. *Journal of Research in Science Teaching*, *47*, 454–473. doi:10.1002/tea.20346

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning*, *3*, 269–276. doi:http://dx.doi.org/10.1016/0149-7189(80)90042-7

Drake, R., Goldman, H., Leff, S., Lehman, A., Dixon, L., Mueser, K., & Torrey, W. (2001). Implementing evidenced-based practices in routine mental health service settings. *Psychiatric Services*, *52*, 179–182.

Dufour, R. (2004). What is a “Professional Learning Community?” *Educational Leadership, 61*(6), 6–11.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–50. doi:10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*, 237–256. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12729182

Dutro, S. (2009). Explicit language instruction. In L. Helma (Ed.), *Literacy Development with English Learners: Research-Based Instruction in Grades K–6* (pp. 40–55). New York, NY: The Guilford Press.

Dutro, S., & Moran, C. (2003). Rethinking English language instruction: An architectural approach. In G. G. Garcia (Ed.), *English learners: Reaching the highest level of English literacy* (pp. 227–265). Newark, DE: International Reading Association.

Dyas, J. V., Togher, F., & Siriwardena, A. N. (2014). Intervention fidelity in primary care complex intervention trials: Qualitative study using telephone interviews of patients and practitioners. *Quality in Primary Care, 22*, 25–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24589148

Echevarria, J., Frey, N., & Fisher, D. (2015). What it takes for English Learners to succeed. *Educational Leadership, 72*(6), 22–26.

Echevarria, J., Richards-Tutor, C., Chinn, V. P., & Ratleff, P. A. (2011). Did they get it? The role of fidelity in teaching English learners. *Journal of Adolescent and Adult Literacy, 54*, 425–434. doi:10.1598/JA

E. L. Achieve (2014). Constructing Meaning Home. Retrieved from http://cm.elachieve.org/

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80,* S46–S54. doi:10.1097/00001888-200510001-00015

Eva, K. W., & Regehr, G. (2008). “I’ll never play professional football” and other fallacies of self-assessment. *The Journal of Continuing Education in the Health Professions, 28*(1), 14–19. doi:10.1002/chp

Fantilli, R. D., & McDougall, D. E. (2009). A study of novice teachers: Challenges and supports in the first years. *Teaching and Teacher Education*, *25,* 814–825. doi:10.1016/j.tate.2009.02.021

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, *92*, 694–702. doi:10.1037/0022-0663.92.4.694

Fink, A. (2013). *How to conduct surveys: A step by step guide* (5th ed.). Thousand Oaks, CA: Sage Publications.

Fives, H., Hamman, D., & Olivarez, A. (2007). Does burnout begin with student-teaching? Analyzing efficacy, burnout, and support during the student-teaching semester. *Teaching and Teacher Education*, *23*, 916–934. doi:10.1016/j.tate.2006.03.013

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy*, *36*, 3–13. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1464400&tool=pmcentrez&rendertype=abstract

Fullan, M. (1991). *The new meaning of educational change* (1st ed.). London, England: Cassell.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2009). What makes professional development effective? Results from a national sample of teachers. *American Education Research Journal*, *38*, 915–945. doi:10.3102/00028312038004915

Gliem, J. A., & Gliem, R. R. (2003). *Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales,. 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education*. doi:10.1109/PROC.1975.9792

Graham, S., & Perin, D. (2007). *Writing Next: Effective strategies to improve writing of adolescents in middle and high schools*. *A Report to Carnegie Corporation of New York*. New York, NY. Retrieved from https://www.paytixx.com/education/nclb/ispd/topic1/writing_next.rtf

Gulumhussein, A. (2013). *Teaching the teachers: Effective professional development in an era of high stakes accountability*. Alexandria, VA. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Teaching-the-Teachers-Effective-Professional-Development-in-an-Era-of-High-Stakes-Accountability/Teaching-the-Teachers-Full-Report.pdf

Guskey, T. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, *8,* 381–391. doi:10.1080/135406002100000512

Hall, G. E., & Hord, S. M. (1987). *Change in schools: Facilitating the process*. New York, NY: State University of New York Press.

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program integrity as a moderator of prevention program effectiveness: Results for fifth-grade students in the adolescent alcohol prevention trial. *Journal of Studies on Alcohol*, *52,* 568–579.

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box : Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology*, *27*, 711–731.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools. *Exceptional Children*, *79*, 181–193.

Hesse-Biber, S., & Johnson, R. B. (2013). Coming at things differently: Future directions of possible engagement with mixed methods research. *Journal of Mixed Methods Research*, *7*, 103–109. doi:10.1177/1558689813483987

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, *53*, 499–503. doi:10.1016/S0895-4356(99)00174-2

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26. doi:10.3102/0013189X033007014

Joyce, B., & Showers, B. (2002). Student achievement through staff development. In B. Joyce & Beverly Showers (Eds.), Designing training and peer coaching: Our needs for learning (pp. 1–5). Alexandria, VA: ASCD.

Kaderavek, J. N., & Justice, L. M. (2010). Fidelity: An essential component of evidence-based practice in speech-language pathology. *American Journal of Speech Language Pathology*, *19*, 369–379.

Knoche, L. L., Sheridan, S. M., Edwards, C. P., & Osborn, A. Q. (2010). Implementation of a relationship-based school readiness intervention: A multidimensional approach to fidelity measurement for early childhood. *Early Childhood Research Quarterly, 25*, 299–313. doi:10.1016/j.ecresq.2009.05.003

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology*, *30*, 242–262. doi:10.1016/j.cedpsych.2004.10.001

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Kwakman, K. (2003). Factors affecting teachers’ participation in professional learning activities. *Teaching and Teacher Education, 19,* 149–170. doi:10.1016/S0742-051X(02)00101-4

Langendyk, V. (2006). Not knowing that they do not know: Self-assessment accuracy of third-year medical students. *Medical Education, 40,* 173–179. doi:10.1111/j.1365-2929.2005.02372.x

Leachman, M., & Mai, C. (2014). *Most states funding schools less than before the recession*. Washington, DC: Center on Budget and Policy Priorities. Retrieved from www.cbpp.org

Learning Forward. (2014). The professional learning associations’s website. Retrieved from www.learningforward.org

Learning Forward. (2014). Standards home page. Retrieved October 10, 2015, from http://learningforward.org/standards#.VhqsUmRViko

Lucero, A. (2013). Teachers’ use of linguistic scaffolding to support the academic language development of first-grade emergent bilingual students. *Journal of Early Childhood Literacy*, *14,* 534–561. doi:10.1177/1468798413512848

Mallette, B., Maheady, L., & Harper, G. F. (1999). The effects of reciprocal peer coaching on preservice general educators’ instruction of students with special learning needs. *Teacher Education and Special Education, 22,* 201–216. doi:10.1177/088840649902200402

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Thousand Oaks, CA: Sage Publications.

McGrew, J. H., Bond, G., Dietzen, L., & Salyers, M. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology*, *62*, 670–678.

McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., Finnerty, M. T. (2007). Fidelity outcomes in the National Implementing Evidence-Based Practices Project. *Psychiatric Services*, *58*, 1279–1284. doi:10.1176/appi.ps.58.10.1279

McKenna, J. W., Flower, A., & Ciullo, S. (2014). Measuring fidelity to improve intervention effectiveness. *Intervention in School and Clinic*, *5*, 1–7. doi:10.1177/1053451214532348

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performancesas scientific inquiry into score meaning.* Princeton, NJ: Educational Testing Service.

Miller, B., Lord, B., & Dorney, J. (1994). *Staff development for teachers: A study of configurations and costs in four districts.* Newton, MA: Education Development Center.

Mitchell, I. (2008). The relationship between teacher behaviours and student talk in promoting quality learning in science classrooms. *Research in Science Education*, *40*, 171–186. doi:10.1007/s11165-008-9106-9

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*, 247–266. doi:10.1016/0272-7358(91)90103-2

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2013). *IBM SPSS for introductory statitics: Uses and interpretations* (5th ed.). New York, NY: Routledge.

Morisky, D. E., Green, L. W., & Levine, D. M. (1986). Concurrent and predictive validity of a self-reported measure of medication adherence. *Medical Care*, *24*, 67–74. doi:10.1097/00005650-198601000-00007

Mortneson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implemntation of a prereferral academic intervention. *School Psychology Review*, *27*, 613–627.

Mowbray, C. T., Bybee, D., Holter, M. C., & Lewandowski, L. (2006). Validation of a fidelity rating instrument for consumer-operated services. *American Journal of Evaluation*, *27*, 9–27. doi:10.1177/1098214005284971

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340. doi:10.1177/109821400302400303

Nishimura, T. (2014). Effective professional development of teachers: A guide to actualizing inclusive schooling. *International Journal of Whole Schooling*, *10*(1), 19–42.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528. doi:10.1111/0023-8333.00136

O'Brien, D. G., Stewart, R. A., & Moje, E. B. (1995). Why content literacy is difficult to infuse into the secondary school: Complexities of curriculum, pedagogy, and school culture. *Reading Research Quarterly, 30*, 442–463.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84. doi:10.3102/0034654307313793

Odden, A., Archibald, S., Fermanich, M., & Gallagher, H. A. (2012). A cost framework for professional development. *Journal of Education Finance*, *28*, 51–74.

Opfer, V. D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*, 376–407. doi:10.3102/0034654311413609

Orwin, R. G. (2000). Assessing program fidelity in substance abuse health services research. *Addiction*, *95,* S309–S327. doi:10.1080/09652140020004250

Ouimet, J. A., Bunnage, J. C., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups, expert advice, and cognitive interviews to establish the validity of a college student survey. *Research in Higher Education*, *45*, 233–250. doi:10.1023/B:RIHE.0000019588.05470.78

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., & Duque, M. (2013). The waive of the future? School accountability in the waiver era. *Educational Researcher*, *43*, 45–54. doi:10.3102/0013189X13517137

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and redommentations. *Research in Nursing & Health*, *29,* 489–497. doi:10.1002/nur

Race to the Top Act of 2011, Pub. L. No. H.R. 1532 (2014).

Rosenshine, B. (1987). Explicit teaching and teacher training. *Journal of Teacher Education*, *38*(3), 34–36. doi:10.1177/002248718703800308

Saad, L. (2014). *U.S. teachers offer split decision on Common Core*. Washington, DC:Gallup Retrieved from http://www.gallup.com/poll/178892/teachers-offer-split-decision-common-core.aspx?version=print

Sankar, A., Golin, C., Simoni, J. M., Luborsky, M., & Pearson, C. (2006). How qualitative methods contribute to understanding combination antiretroviral therapy adherence. *Journal of Acquired Immune Deficiency Syndromes (1999)*, *43 Suppl 1*, S54–S68. doi:10.1097/01.qai.0000248341.28309.79

Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education, 27,* 396–407. doi:10.1177/088840640402700407

Sheen, R. (2002). "Focus on form" and "focus on forms." *English Language Teaching*, 56, 303–305.

Shymansky, J. A., Wang, T.-L., Annetta, L. A., Yore, L. D., & Everett, S. A. (2010). How much professional development is needed to effect positive gains in K–6 student achievement on high stakes science tests? *International Journal of Science and Mathematics Education, 10*, 1–19. doi:10.1007/s10763-010-9265-9

Singh, S., & Fletcher, K. E. (2014). A qualitative evaluation of geographical localization of hospitalists: How unintended consequences may impact quality. *Journal of General Internal Medicine*, *29*, 1009–1016. doi:10.1007/s11606-014-2780-6

Skaalvik, E. M., & Skaalvik, S. (2010). Teacher self-efficacy and teacher burnout: A study of relations. *Teaching and Teacher Education*, *26*, 1059–1069. doi:10.1016/j.tate.2009.11.001

Slabine, N. A. (2011). *Evidence of effectivness*. Oxford, OH: Learning Forward.

Sugai, G., & Horner, R. H. (2002). Introduction to the special series on Positive Behavior Support in schools. *Journal of Emotional and Behavioral Disorders*, *10*, 130–135. doi:10.1177/10634266020100030101

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*, 963–980. doi:10.1002/1098-2736(200011)37:9<963::AID-TEA6>3.0.CO;2-0

Tan, L., Sun, X., & Khoo, S. T. (2014). Can engagement be compared? Measuring academic engagement for comparison. In *International Conference on Educational Data Mining* (pp. 213–216). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/short papers/213_EDM-2014-Short.pdf

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2,* 53–55. doi:10.5116/ijme.4dfb.8dfd

Thorndike, R. M., & Throndike-Christ, T. (2011). *Measurement and Evaluation in Psychology and Education* (8th ed.). Upper Saddle River, NJ: Pearson Education.

Tompkins, G., Campbell, R., Green, D., & Smith, C. (2014). *Literacy for the 21st century* (2nd ed.). Melborne: Pearson Australia.

Torrey, W. C., Lynde, D. W., & Gorman, P. (2005). Promoting the implementation of practices that are supported by research: The national implementing evidence-based practice project. *Child and Adolescent Psychiatric Clinics of North America, 14,* 297–306.

U.S. Department of Education. (2012). *ESEA Flexibility*. Washington, DC.

Vacca, R. T., & Vacca, J. A. L. (1989). *Content area reading*. Glenview, IL: Scott, Foresman.

Vaden-Kiernan, M., Jones, D. H., & McCann, E. (2009). *Latest eveidence on the National Staff Development Council's Standards Assessment Inventory*. Austin, TX.

Valdés, G., Kibler, A., & Walqui, A. (2014). *Changes in the expertise of ESL professionals: Knowledge and action in an era of new standards*. Alexandria, VA. Teachers of English to Speakers of Other Languages (TESOL).

Wagner, B. D., & French, L. (2010). Motivation, work satisfaction, and teacher change among early childhood teachers. *Journal of Research in Childhood Education, 24*, 152–171. doi:10.1080/02568541003635268

Warren, C. A. B. (2002). Qualitative interviewing. In J. Gubrium, J. Holstein, *Handbook of interview research: Context and method* (pp. 230–258). Thousand Oaks, CA: Sage Publications.

Webster-Stratton, C., Reinke, W. M., Herman, K. C., & Newcomer, L. L. (2011). The incredible years teacher classroom management training: The methods and principles that support fidelity of training delivery. *School Psychology Review*, *40,* 509–529.

Weiss, M. J., Bloom, H. S., & Brock, T. (2013, June). A conceptual framework for studying the sources of variation in program effects. *MDRC Working Papers on Research Methodology*. Retrieved from http://mdrc.org/sites/default/files/a-conceptual_framework_for_studying_the_sources.pdf

Wolery, M. (2011). Intervention research: The importance of fidelity measurement. *Topics in Early Childhood Special Education*, *31*, 155–157. doi:10.1177/0271121411408621

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*, 156–67. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7217482

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation*, *30*, 44–61. doi:10.1177/1098214008329523

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation*, *33*, 547–565. doi:10.1177/1098214012452715